UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MIARFID Official Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

Department of Computer Systems and Computing
Polytechnic university of Valencia

Search and information extraction in handwritten tables

Master Thesis

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

Author: José Andrés Moreno

Tutor: Enrique Vidal Ruiz

Course 2020-2021

Acknowledgements

I wanted to thank Enrique, for always finding a moment to help me when I have needed it and for being open to discuss all the issues I have raised during this thesis.

Resum

Actualment, arxius de tot el món estan digitalitzant grans col·leccions de documents manuscrits amb la finalitat de preservar-los i facilitar la seua difusió a investigadors i usuaris generals. Aquest fet està motivant una gran evolució en les tècniques de reconeixement de text manuscrit (HTR per les seues sigles en anglés), que permeten accedir als continguts textuals de les imatges digitals mitjançant consultes de text pla, de la mateixa manera que es fa amb els llibres i altres documents digitals.

Dins del conjunt de documents manuscrits sense transcripció, ens trobem que aproximadament més de la meitat dels documents es corresponen amb documents estructurats. Aquests documents contenen informació de tota mena: registres de naixement, de navegació, quaderns de bitàcola, etc. Tota aquesta informació és sovint imprescindible per a usos jurídics, estudis demogràfics, estudis de l'evolució del clima, etc.

L'objectiu d'aquest treball és desenvolupar nous mètodes que permeten realitzar cerques segons el model "atribut-valor" sobre aquests documents, on els "atributs" són les capçaleres de les columnes i files que formen la taula i els "valors" són la resta de cel·les de la taula que no són capçalera. Per a això, ens basarem en el marc de la indexació probabilística (que està en certa manera relacionat amb el camp conegut com "keyword spotting"). En aquest marc, cada element d'una imatge que es puga interpretar com una paraula és detectat i emmagatzemat, juntament amb la seua posició dins de la imatge i la corresponent probabilitat de rellevància.

Així doncs, emprant la informació geomètrica dels índexs probabilístics en conjunt amb l'ús de distribucions Gaussianes, es pretén permetre realitzar aquest tipus de cerques des d'una perspectiva completament probabilística. Sota aquest enfocament, a més de la cerca, s'estudia l'extracció de la informació amb objectiu de bolcar els continguts específics de les imatges digitals a un format compatible amb bases de dades convencionals. En totes dues tasques s'han aconseguit resultats que superen el baseline proposat.

Paraules clau: Reconeixement de Formes, Processat d'Imatges, Documents Estructurats Manuscrits, Indexació Probabilística i Cerca, Extracció d'Informació

Resumen

Actualmente, archivos de todo el mundo están digitalizando grandes colecciones de documentos manuscritos con el fin de preservarlos y facilitar su difusión a investigadores y usuarios generales. Este hecho está motivando una gran evolución en las técnicas de reconocimiento de texto manuscrito (HTR por sus siglas en inglés), que permiten acceder a los contenidos textuales de las imágenes digitales mediante consultas de texto plano, de la misma manera que se hace con los libros y otros documentos digitales.

Dentro del conjunto de documentos manuscritos sin transcripción, nos encontramos con que aproximadamente más de la mitad de los documentos se corresponden con documentos estructurados. Estos documentos contienen información de todo tipo: registros de nacimiento, de navegación, cuadernos de bitácora, etc. Toda esta información es a menudo imprescindible para usos jurídicos, estudios demográficos, estudios de la evolución del clima, etc.

El objetivo de este trabajo es desarrollar nuevos métodos que permitan realizar búsquedas según el modelo “atributo-valor” sobre estos documentos, donde los “atributos” son las cabeceras de las columnas y filas que forman la tabla y los “valores” son el resto de celdas de la tabla que no son cabecera. Para ello, vamos a basarnos en el marco de la indexación probabilistica (que está en cierto modo relacionado con el campo conocido como “keyword spotting”). En este marco, cada elemento de una imagen que se pueda interpretar como una palabra es detectado y almacenado, junto con su posición dentro de la imagen y la correspondiente probabilidad de relevancia.

Así pues, empleando la información geométrica de los índices probabilísticos en conjunto con el uso de distribuciones gausianas, se pretende permitir realizar este tipo de búsquedas desde una perspectiva completamente probabilística. Bajo este enfoque, además de la búsqueda, se estudia la extracción de la información con objetivo de volcar contenidos específicos de las imágenes digitales a un formato compatible con bases de datos convencionales. En ambas tareas se han logrado resultados que superan el baseline propuesto.

Palabras clave: Reconocimiento de Formas, Procesado de Imágenes, Documentos Estructurados Manuscritos, Indexación Probabilística y Búsqueda, Extracción de Información

Abstract

Currently, all archives around the world are digitising large collections of manuscripts, aiming to preserve and facilitate their dissemination to researchers and general users. This fact is motivating a fast evolution in handwritten text recognition (HTR) techniques, which allow accessing to the textual contents of digital images by means of plain-text queries, in the same way as with books and other digital documents.

Among the huge set of manuscripts without transcription, more than half of the documents contain structured text. This is the case of birth records, navigation logs, etc. The information contained in these documents is often needed for legal matters, demographic studies, weather evolution studies, etc.

The purpose of this work is to develop new methods that allow to perform searches according to the "attribute-value" model about these documents, where the "attributes" are, for example, column or row headers in tables and the "values" are the corresponding table cells. For this purpose, we will rely on the so-called probabilistic indexing framework (which in a certain sense is related with the field known as "keyword spotting"). In this framework, each element of an image that can be interpreted as a word is detected and stored, along with its position within the image and the correspondence relevance probability.

This way, by using the geometric information available in the probabilistic indices and Gaussian distributions, we aim at allowing this type of search from a completely probabilistic perspective. Following this approach, in addition to information search, we study how to actually extract specific textual contents of the digital images in standard formats compatible with conventional databases.

Key words: Pattern Recognition, Image Processing, Structured Handwritten Documents, Probabilistic Indexing and Search, Information Extraction

Contents

List of Figures

List of Tables

CHAPTER 1
Introduction

Nowadays, archives from all over the world are digitising huge collections of handwritten documents without transcription with the purpose of preserving them and facilitating their dissemination among researchers and users in general. However, despite digitising documents helps to preserve them, their textual richness is still hidden behind the billions of pixels that conform these images. This fact is motivating an evolution on the Handwritten Text Recognition techniques (HTR), which allows accessing to the textual content of the images through plain text, in the same way as it is done in digital books.

1.1 Motivation

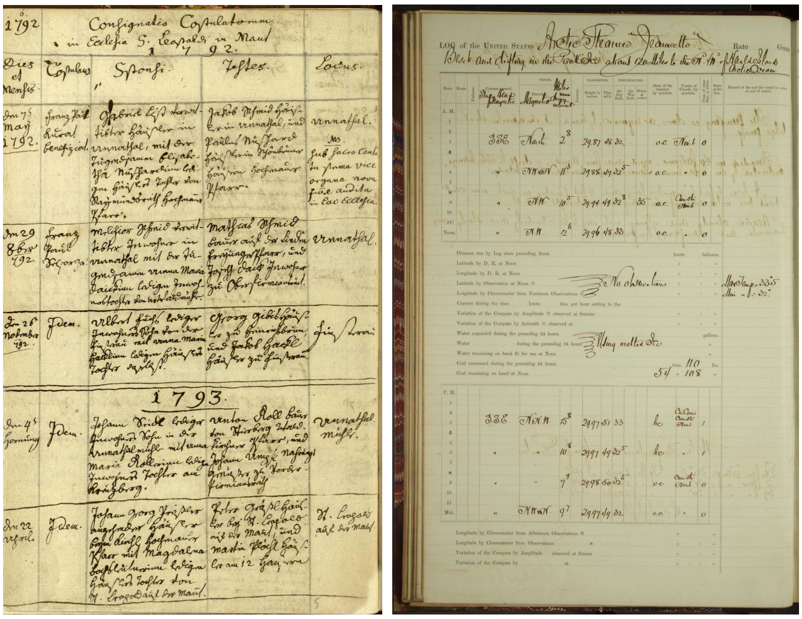
Within the set of handwritten documents without transcription, we find that approximately more than half of them correspond to structured documents. These documents contain all kinds of information: birth, marriage and death registers, notarial data, navigation records, logbooks, etc. All this information is often essential with juridical purposes, to perform demographic and genealogical studies, analyse the weather's evolution, etc. Given this information, it becomes a relevant task to obtain a reliable automatic transcription of their contents.

This task is very challenging due to a variety of factors:

In first place, the layout of the tables might be variable, inconsistent and even erratic, given that depending on the collection of documents that we are considering, the tables can be completely handwritten (structure and contents), hybrid (printed structure and handwritten contents) or even completely printed (structure and contents). Examples are shown in Fig. 1.1.

In second place, the text lines that conform the contents of the table are generally shorter than the lines present in regular handwritten documents. This is due to the fact that they typically account for concrete and short attributes such as proper names, ages, coordinates, sailing data, etc. This fact makes the line detection more difficult, but also the word recognition as the shorter lines lack the linguistic context, which typically helps to provide accurate hypotheses.

In addition to these difficulties, we would like to remark that we are looking for not only an automatic transcription, but one that takes into account the nature and relationships found in the textual contents of the documents in order to unleash all



1792 Consignatio Copulatorum in Ecclesia S. Leopoldi in Maut 1792

Dies et Mensis	Copulans	Sponsi	Testes	Locus

1793

LOG of the United States Arctic Steamer Jeannette

Figure 1.1: Example of two documents containing tables, the first one corresponds to the Passau dataset meanwhile, the second one corresponds to the HisClima dataset. In the first table, it can be seen that the structure and contents of the image are handwritten, whereas in the second image, the structure of the table is printed and its contents are handwritten.

the possibilities that these documents have to offer. Therefore, the main purpose of this thesis is to allow users to perform tabular queries, which are queries where the user specifies the column and/or row that he wants to query, as well as the value that he is looking for in that column and/or row. Moreover, given that many times it is also necessary to extract all this information to fill an external database, we also consider the information extraction task.

1.2 State of the art

The information search and extraction over structured handwritten documents is a relevant task that has been attacked over the years with different perspectives and techniques. Typically, it has been solved by first performing Layout Analysis over the images to infer information about the structure of the images (such as the lines that conform a table, the box that denotes a register, etc.) and then, perform HTR over the delimited zones.

Following this approach, in [2] they first recognise the graphical lines that conform the structure of the table. Then, they perform template matching over the graphical structure and obtain the columns and the headers structure of the table. Finally, they use two different approaches (Conditional Random Fields and Graph Convolutional Networks) to determine which column and row correspond to each cell.

In [8], they firstly locate the table on the page employing an algorithm that makes use of the printed anchors of the table. Once they have located the table, they split the table into the different rows and columns that conform it. Finally, they apply a CNN or an RNN to transcribe the textual content of each cell.

Despite being possible to face this task as two different consecutive problems (the detection of the cell region and the transcription of the cell contents), other authors have proposed layout-agnostic solutions, where the table's structure is directly inferred from the geometric information of the transcription of the contents.

In [7] there is proposed an information search and extraction technique that is layout-agnostic. In this technique, the column headers of the tables are considered "anchors" that determine the width of the queried column. Therefore, after applying geometric reasoning over the document, it can be inferred the structure of the table and it is straightforward to allow the users to perform tabular queries.

The major drawback of this approach is that despite having achieved good results over a variety of different collections [7, 11], it relies on holistic assumptions to determine, during the geometric reasoning step, the width of the columns that conform the tables. In this thesis, we aim at substituting these heuristics by well-known statistical models, which are more robust and theoretically motivated.

1.3 Objectives

Therefore, the main objective of this thesis is to allow the users to perform tabular queries over the images of the handwritten manuscripts, employing a probabilistic framework that replaces all the heuristics that are found in [7] by well-known statistical models. With this purpose, we have defined the following minor objectives:

- Propose a probabilistic framework for tabular queries.
- Try different machine learning models to estimate the required probabilities.
- Measure the performance of the proposed systems when searching and extracting information.

1.4 Thesis structure

In order to explain how to perform information search and extraction over these documents, this thesis is structured in 5 chapters:

In Chapter 2 we explain the theoretical foundations that are employed in this work, as well as the performance metrics employed.

In Chapter 3 we detail our proposal of probabilistic framework to perform different types of tabular queries, as well as an explanation on how we have adapted this framework to perform information extraction.

In Chapter 4 can be found the experimental results obtained when performing search and information extraction, along with an exploratory analysis of the data and a description of the employed corpus.

Finally, in Chapter 5 we conclude and propose possible future works that could be explored after this thesis.

CHAPTER 2
Theoretical foundations

This section details the theoretical foundations considered in this thesis, giving a brief explanation of their nature and use. Moreover, we provide an explanation of the different evaluation measures considered.

2.1 Technological context

2.1.1. Graphical Models

Graphical models [1] can be defined as a compact and graphical representation of the joint distribution of a set of variables employing directed graphs (Bayesian networks) or undirected graphs (Markov random fields). This representation is very powerful, as it combines graph and probability theory elegantly into a unique model. Among the different useful properties that these networks have, we want to highlight the following:

Firstly, they provide an intuitive and easy way to visualise and create new probabilistic models. Secondly, they provide details about the structure of the model, such as the conditional independence between variables, their relationships, etc. Finally, they provide a mechanical way of inferring and learning complex models through graphical manipulations.

Now, we are going to focus on Bayesian networks because we have employed them in this work.

A Bayesian network is defined as a Directed Acyclic Graph (DAG) where nodes represent random variables and edges represent the different dependencies between the different variables that conform the network. A Bayesian network defines a joint probability distribution for nodes $x_1, x_2, ..., x_n$ as:

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{D} P(x_i \mid a(x_i))$$

where $a(x_i)$ denotes the values of the variables associated to the ancestors of x_i.

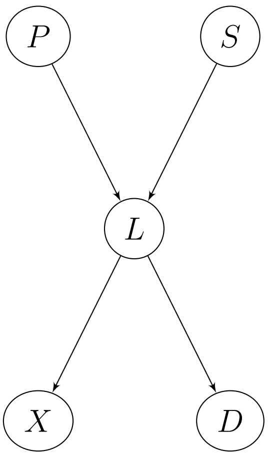


Figure 2.1: Example of Bayesian network.

Fig. 2.1 shows an example of a simple bayesian network, which models the relationship between symptoms and external factors that are related to lung cancer. The random variables considered are the following:

- P: Denotes the degree of external pollution received by the person during his life, either high (h) or low (l).
- S: Denotes if the patient is a smoker, either yes (y) or no (n).
- D: Denotes if the patient suffers dyspnoea, either yes (y) or no (n).
- X: Denotes the result of the x-ray radiography. It can be either positive (p), uncertain (u) or negative (n).
- L: Denotes if the patient has lung cancer, either yes (y) or no (n).

Then, following the network defined in Fig. 2.1, the joint probability can be calculated:

$$P(P,S,L,D,X) = P(P)\,P(S)\,P(L \mid P,S)\,P(X \mid L)\,P(D \mid L)$$

Please note that independence assumptions are made explicit graphically in the network. For instance, the fact that the degree of external pollution received by a person is conditionally independent of the fact that he is a smoker or not.

Now, employing the Bayes theorem, we can make inferences easily employing the network structure. For example: Which is the probability that a person does not suffers lung cancer given that he does not smoke, the results of the X-ray are negative but he suffers dyspnoea? This question can be modelled as:

$$\begin{aligned} P(L=n \mid P,S=n,D=y,X=n) &= \frac{\sum_{i\epsilon P} P(L=n,P=i,S=n,D=y,X=n)}{\sum_{i\epsilon P} P(P=i,S=n,D=y,X=n)} \\ &= P(L=n \mid P,S=n) \end{aligned}$$

2.1.2. Gaussian distribution

In order to model the different probabilities that conform our probabilistic framework, we have assumed that the different variables of our model follow a Gaussian distribution.

The Gaussian distribution is a parametric model that estimates the probability distribution $p(\mathbf{x})$ of a random variable $\mathbf{x}$, given a finite set of observations $\mathbf{x}_1$, $\mathbf{x}_2$,..., $\mathbf{x}_N$. This model is governed by two parameters, the mean, which denotes the centre of the probability distribution, and the variance, which denotes the "spread" of the distribution.

We start by defining the particular case of the univariate Gaussian, which is defined by the following formula:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$

where x is the single variable we want to estimate its probability, μ is the mean value and σ^2 is the variance. Examples of univariate Gaussian distributions with different values for their parameters can be found in Fig. 2.2. It can be seen explicitly that the centre of each distribution corresponds to the value of μ and that the probability mass is distributed according to σ^2 (the larger σ^2, the spreader the associated probability distribution).

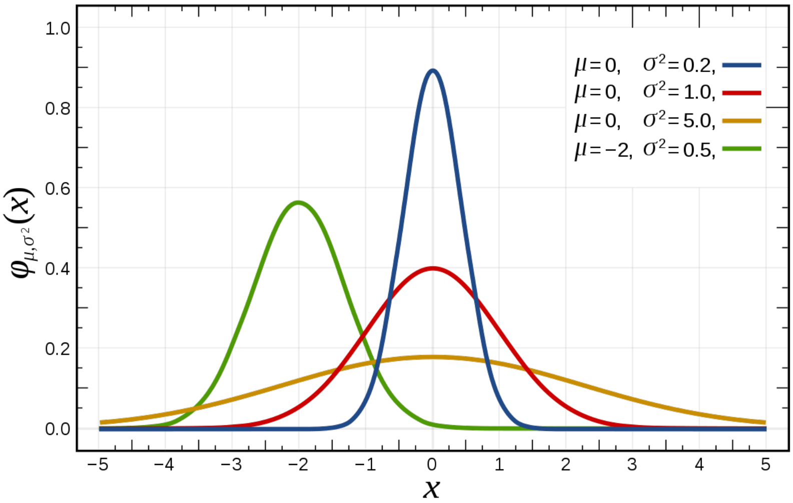


Figure 2.2: Different Gaussian distributions according to the values of the parameters μ and σ^2. Extracted from https://en.wikipedia.org/wiki/Normal_distribution the 21^{st} of June 2021.

Now, the formula that defines the multivariate Gaussian is:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu}) \right\}$$

where $\mathbf{x}$ denotes the D-dimensional vector of which we want to estimate its probability, $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. It is worth noting that the number of free parameters in $\boldsymbol{\Sigma}$ is $D(D+1)/2$. Given this, the number

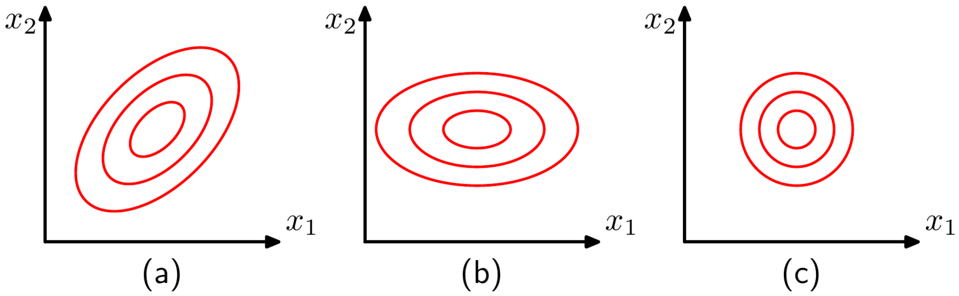


Figure 2.3: Shape of three different gaussians according to $\mathbf{\Sigma}$. a) Accounts for a full covariance matrix, b) accounts for a diagonal covariance matrix and c) accounts for a covariance matrix proportional to the identity matrix. Extracted from *Pattern recognition and Machine learning*, (2006) Bishop, C. M.

of parameters grows quadratically with D and, for large values of D, it might be prohibitive to operate and manipulate $\mathbf{\Sigma}$. As a possible solution to this problem, the covariance matrix can be constrained to be diagonal, reducing the number of parameters to D, or even to be proportional to the identity matrix. In the case of constraining $\mathbf{\Sigma}$ to be diagonal, we will find that the isometric contours that define the distribution are aligned with the coordinate axis, meanwhile in the case of being proportional to the identity matrix, the contours are concentric circles. Moreover, we would like to remark that these constraints might be interesting not only to reduce the number of parameters to estimate but also to avoid overfitting. Examples of Gaussian distributions in $2D$ with different $\mathbf{\Sigma}$ can be found in Fig. 2.3.

Finally, in order to estimate the parameters that govern each Gaussian, we have employed the maximum likelihood criterion. The formulas to estimate them are the following:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n} \mathbf{x_n}$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{N} \sum_{n} (\mathbf{x_n} - \hat{\boldsymbol{\mu}})(\mathbf{x_n} - \hat{\boldsymbol{\mu}})^{\mathrm{T}}$$

2.1.3. Probabilistic Indexing framework

In order to deal with the intrinsic word-level uncertainty generally exhibited by handwritten text in images and in particular handwritten structured documents, as discussed in Chap. 1, we employ the Probabilistic Indexing [16] (PrIx) framework. This framework draws from ideas and concepts previously developed for keyword spotting (KWS), both in speech signals and text images. However, rather than caring for "key" words, any element in an image which is likely enough to be interpreted as a word is detected and stored, along with its relevance probability (RP) and its location in the image. These text elements are referred to as "*pseudo-word spots*".

KWS can be seen in this context as the binary classification problem of deciding whether a particular image region x is relevant for a given query word v, i.e. try to answer the following question: "Is v actually written in x?". As in [15, 9, 16], we denote this image-region word RP as $P(R = 1 \mid X = x, V = v)$, but for the sake of conciseness, we will omit the random variable names, and for $R = 1$, we will simply write R.

As discussed in [16], this RP can be simply approximated by the maximum value of the *posteriorgram* for v in x, $P(v \mid x, i, j)$, which denotes the probability that v is written in a subimage of x which includes the pixel (i, j). An example a of full *posteriorgram* for $P(v = \text{“matter”} \mid x, i, j)$ can be found in Fig. 2.4. Therefore:

$$P(R \mid x, v) \;\approx\; \max_{i,j \sqsubseteq x} P(v \mid x, i, j) \tag{2.1}$$

where $\sqsubseteq$ denotes geometric containment (i.e., $i, j \sqsubseteq x$ is the set of coordinates of pixels contained in x).

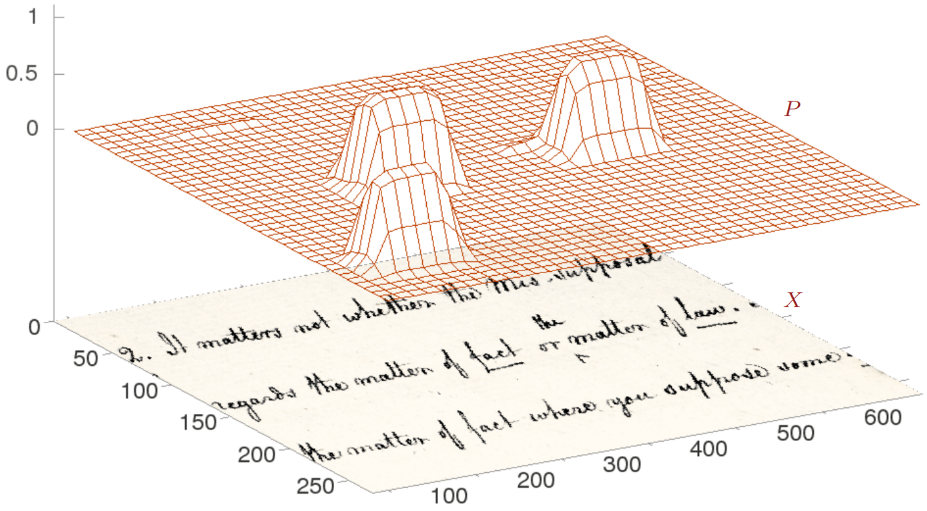


Figure 2.4: Example of full posteriorgram at pixel level for $P(v = \text{“matter”} \mid x, i, j)$

This expression can be more conveniently written in terms of all possible small, word-sized image sub-regions or bounding boxes (BB) in x which may contain a depiction of the writing of the word v as:

$$P(R \mid x, v) \;\approx\; \max_{b \sqsubseteq x} \max_{i,j \sqsubseteq b} P(v \mid x, b, i, j) \;\approx\; \max_{b \sqsubseteq x} P(v \mid x, b) \tag{2.2}$$

where $P(v \mid x, b)$ is the posterior probability needed to “recognise” the BB image (x, b). Therefore, assuming the computational complexity entailed by the maximisation in Eq. (2.2) is algorithmically managed, any sufficiently accurate isolated word classifier can be used to obtain $P(R \mid x, v)$. Image region word RPs do not explicitly take into account where the considered words may appear in the region x, but the precise positions of the words within x are easily obtained as a by-product. According to Eq. (2.2), the best BB for v in the image region x can be obtained as:

$$\hat{b}_v = \arg\max_{b \sqsubseteq x} P(v \mid x, b) \tag{2.3}$$

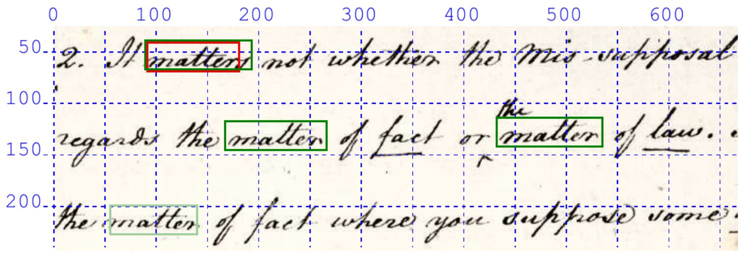
If image regions are small (for example text-line regions), it is unlikely that interesting words appear more than once in the region. Therefore the BB obtained in Eq. (2.3) generally provides good information about the location and size of v within x. This can be straightforwardly extended to cases where the image regions are larger (for example page images) and/or if multiple instances of the same interesting word are expected to appear in x. To do this, rather than finding the (single) best BB in Eq. (2.3), the n-best BBs and the corresponding relevance probabilities can be obtained as:

$$\hat{b}_1, ..., \hat{b}_n = n\text{-}\underset{b \sqsubseteq x}{\text{best}}\ P(v \mid x, b), \quad P_i(\mathcal{R} \mid x, \hat{b}_i) = P(v \mid x, \hat{b}_i),\ 1 \leq i \leq n \tag{2.4}$$

By setting n large enough, all the sufficiently relevant location and size hypothesis for a word v in x are obtained. As a result, the PrIx of a image x consists of a list of *spots* of the form:

$$[x, v, P_i, \hat{b}_i], \ \ P_i \overset{\text{def}}{=} P(v \mid x, \hat{b}_i), \quad v \in V,\ 1 \leq i \leq n \tag{2.5}$$

where V is a set of relevant words or "*pseudo-words*". An example of a PrIx can be found in Fig. 2.5


0
100
200
300
400
500
600
50
100
150
200
2. It matters not whether the mis-supposal
regards the matter of fact or the the matter of law.
the matter of fact where you suppose some

```
# pageID="Bentham-071-021-002-part"
#    keyword confid     bounding box
#
           2  0.929    1   36   20 31
          21  0.064    1   36   24 31
          IT  0.982   33   36   27 31
          IF  0.012   33   36   26 31
     MATTERS  0.989   77   36   99 31
      MATTER  0.011   77   36   93 31
         NOT  0.999  216   36    7 31
     WHETHER  1.000  256   36   99 31
         THE  0.997  389   36   33 31
MIS-SUPPOSAL  1.000  455   36  193 31

         THE  0.927  430   88   30 31
          HE  0.056  434   88   25 31
         ...  ...      ...      ...
```

```
REGARDS  0.857    5 115   84 31
REWARDS  0.138    5 115   90 31
    THE  0.993  110 115   43 31
 MATTER  0.998  160 115   93 31
     OF  0.996  271 115   23 31
   FACT  0.999  306 115   49 31
     OR  0.973  377 115   37 31
     ON  0.021  377 115   42 31
 MATTER  0.990  425 116  100 31
     OF  0.995  542 115   25 31
     BY  0.407  575 115   30 31
    ANY  0.175  575 115   55 31
    ...  ...      ...      ...
    LAW  0.032  575 115   36 31
    LAY  0.031  575 115   55 31
    ...  ...      ...      ...
    PAY  0.012  575 115   59 31
```

```
     THE  0.990    1 198   28 31
  MATTER  0.934   61 198   64 31
      OF  0.988  141 198   28 31
    FAST  0.367  182 198   62 31
     FAR  0.186  182 198   36 31
     ...  ...      ...      ...
    FACT  0.017  182 198   46 31
      AS  0.142  200 198   29 31
     HAS  0.022  200 198   29 31
   WHERE  0.992  255 198   90 31
     YOU  0.761  365 198   45 31
    YOUR  0.030  365 198   47 31
    GOES  0.064  372 198   45 31
 SUPPOSE  0.975  429 198  120 31
SUPPOSED  0.024  429 198  125 31
    SOME  0.834  570 198   78 31
  SOONER  0.016  576 198   83 31
     ONE  0.109  580 198   65 31
      ME  0.022  620 198   22 31
```

Figure 2.5: Example of PrIx. Firstly, we can observe the ID of the page at the beginning of the index. Then, we find the data stored in a tabular format. In each row of the table, we find the stored *pseudo-word*, the relevance probability for that pseudo-word in that position and the corresponding BB. Moreover, in this image we find marked the spots for "matter" and "matters" according to their probability.

2.2 Performance metrics

As performance metrics for this work, we have considered five well-known information retrieval metrics: the precision, the recall, the Average Precision, the Mean Average Precision and the F-measure.

Firstly, let $\mathcal{Q}$ be a set of queries and let τ be a specified threshold. Then, we define the recall, $\rho(q, \tau)$ and the precision, $\pi(q, \tau)$, for a given query $q \epsilon \mathcal{Q}$ as:

$$\rho(q, \tau) = \frac{h(q, \tau)}{r(q)}, \qquad \pi(q, \tau) = \frac{h(q, \tau)}{d(q, \tau)}$$

where $r(q)$ denotes the number of relevant image regions to the query q according to the ground-truth, $h(q, \tau)$ denotes the number of relevant regions, according to the ground-truth, retrieved by our system when querying q and $d(q, \tau)$ accounts for the number of retrieved images when querying for q. It can be seen that the precision denotes the proportion of matches that is relevant among all the matches retrieved by the system, meanwhile the recall denotes the proportion of relevant matches found in the ground truth that is retrieved by the evaluated system.

Typically, the use of these performance metrics shows an interesting trade-off that is present when performing information search and extraction. For instance, if a user chooses a high value for the threshold τ, it is likely that the system's precision will improve, as the higher threshold will reduce the number of false positives. However, it is also likely that the system's recall is going to decrease, as some of the relevant regions might have a lower confidence score associated than the minimum threshold, increasing then the number of false negatives. Thus, if instead of increasing the threshold value, the user would have decided to decrease it, we would have probably observed the opposite phenomenon, an increase in the recall and a decrease in the system's precision. This trade-off can be seen graphical through R-P curves [4], where for each threshold, the precision and recall are plotted. An example of R-P curve can be seen in Fig. 2.6.

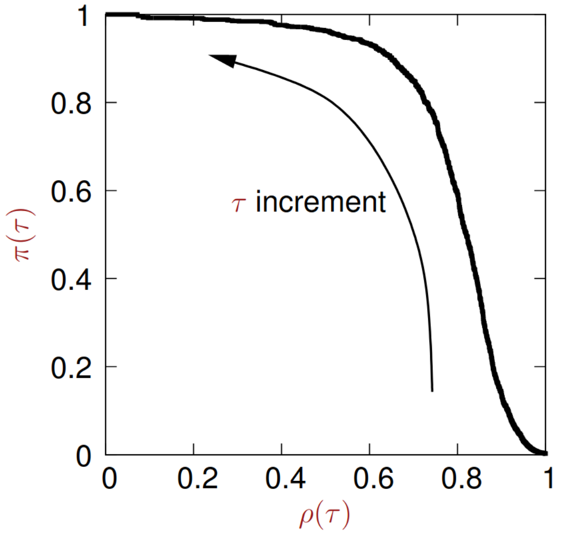


Figure 2.6: Example of R-P curve. It can be seen that, as the threshold increments, typically the precision increases and the recall decreases.

A well-known measure that takes into account this trade-off is the Average Precision (AP) [10], which is defined as the area under the R-P curve. Formally, it is defined as:

$$\bar{\pi}_q = \int_0^1 \pi_q(\rho) d\rho$$

Another well-known information retrieval metric that takes into account this trade-off is the Mean Average Precision (mAP), which is the arithmetic average of the AP of all the queries that form $\mathcal{Q}$. Formally, it is defined as:

$$\bar{\bar{\pi}}_q = \frac{1}{|\mathcal{Q}|} \sum_{q \epsilon \mathcal{Q}} \bar{\pi}_q \tag{2.6}$$

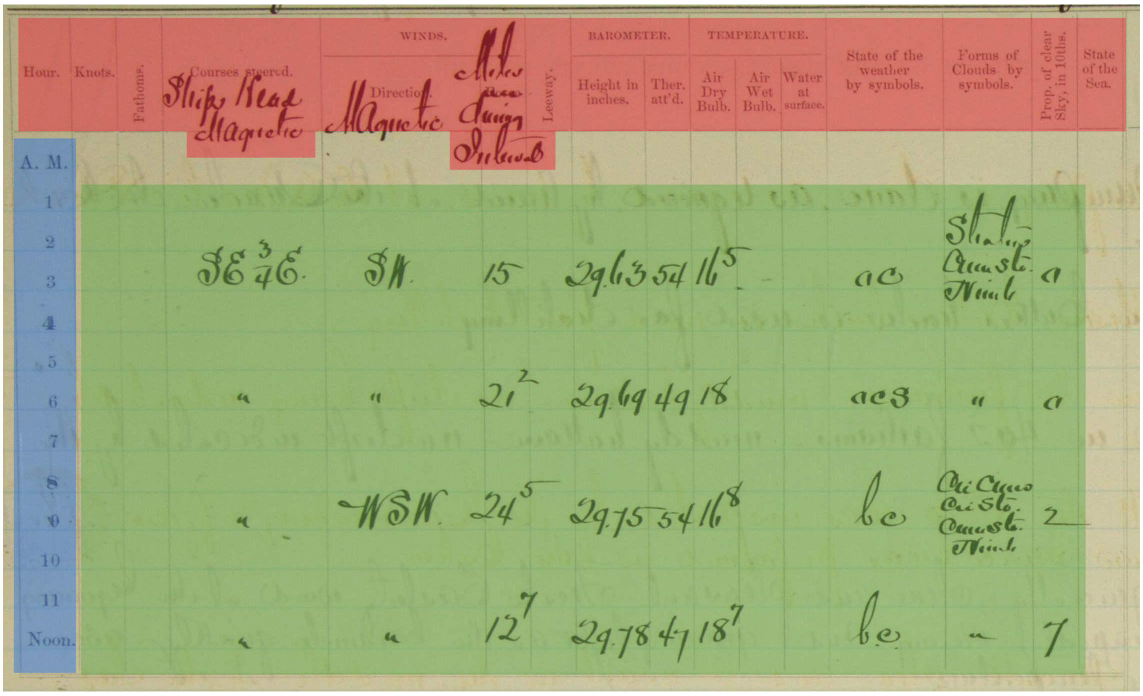
Finally, the F_1-measure is an information retrieval metric that takes into account the balance between precision and recall. Formally, it is defined as:

$$\mathrm{F}_1(q, \tau) = 2 * \frac{\pi(q, \tau) * \rho(q, \tau)}{\pi(q, \tau) + \rho(q, \tau)}$$

CHAPTER 3
Proposed probabilistic framework

In this section, we present our probabilistic framework to perform column queries and row queries. Moreover, we also discuss how to perform the intersection of these two types of queries and how to adapt this framework to perform information extraction.

In our proposed framework, we wish to be agnostic with respect to possibly predefined table layouts, but we assume tables to be organised into orthogonal rows and columns. Each column typically has a column header and each row may have (or not) a row header. The textual content of interest is contained in value cells which are the intersection of columns and rows. An example of table can be seen in Fig. 3.1.



Hour.	Knots.	Fathoms.	Courses steered.	WINDS.			BAROMETER.		TEMPERATURE.			State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
			Ship Head Magnetic	Direction. Magnetic	Miles Force During Interval	Leeway.	Height in inches.	Ther. att'd.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.				
A. M.															
1															
2															
3			SE$\frac{3}{4}$E.	SW.	15		29.63	54	16^{5}			ac	Strat. Cumst. Nimb	a	
4															
5															
6			"	"	21^{2}		29.69	49	18			acs	"	a	
7															
8															
9			"	WSW.	24^{5}		29.75	54	16^{8}			bc	Ci Cum Ci Str. Cumst. Nimb	2	
10															
11															
Noon.			"	"	12^{7}		29.78	47	18^{7}			bc	"	7	

Figure 3.1: Example of table of the HisClima dataset. It is denoted in red the region of the table that corresponds to column headers, in blue the region that corresponds to row headers and in green the region that corresponds to value cells.

3.1 Col-header Cell-Value Queries

Firstly, we consider the problem of performing column queries. We want to compute the relevance probability $P(\mathcal{R} \mid x, q)$, where x is a table image and q is a query. In this subsection we assume $q = (q_{ch}, q_v)$, where q_{ch} is a column header "attribute" query and q_v a value or cell query. For simplicity, we assume that the regions of x matching q_{ch} and q_v are small bounding boxes (BBs), b_{ch}, b_v , respectively, each tightly containing the words specified by the corresponding query.

To simplify notation, we will drop x from the formulation, assuming that all the probabilities are always conditioned by x. Therefore:

$$P(\mathcal{R} \mid x, q) \equiv P(\mathcal{R} \mid q) \equiv P(\mathcal{R} \mid q_{ch}, q_v)$$

$P(\mathcal{R} \mid x, q)$ can be approximated as:

$$\begin{aligned} P(\mathcal{R} \mid x, q) &= \sum_{b_{ch}, b_v \sqsubseteq x} P(\mathcal{R}, b_{ch}, b_v \mid q_{ch}, q_v) \approx \max_{b_{ch}, b_v \sqsubseteq x} P(\mathcal{R}, b_{ch}, b_v \mid q_{ch}, q_v) \\ &= \max_{b_{ch}, b_v \sqsubseteq x} P(b_{ch} \mid q_{ch}, q_v)\, P(b_v \mid b_{ch}, q_{ch}, q_v)\, P(\mathcal{R} \mid b_{ch}, b_v, q_{ch}, q_v) \\ &\approx \max_{b_{ch}, b_v \sqsubseteq x} P(b_{ch} \mid q_{ch})\, P(b_v \mid b_{ch}, q_v)\, P(\mathcal{R} \mid b_{ch}, b_v, q_{ch}, q_v) \\ &\approx \max_{b_{ch}, b_v \sqsubseteq x} P(b_{ch})\, P(b_v \mid b_{ch})\, P(\mathcal{R} \mid b_{ch}, b_v, q_{ch}, q_v) \end{aligned} \tag{3.1}$$

where, as in [16], $P(\mathcal{R} \mid b_{ch}, b_v, q_{ch}, q_v)$ can in turn be approximated as:

$$P(\mathcal{R} \mid b_{ch}, b_v, q_{ch}, q_v) \approx P(q_{ch}, q_v \mid b_{ch}, b_v) \approx \min(P(q_{ch} \mid b_{ch}),\ P(q_v \mid b_v)) \tag{3.2}$$

Firstly, we approximate the sum over all the possible BBs for b_{ch} and b_v by the maximum, as has been done in Eq. (2.1). Then, we consider q_{ch} and q_v conditionally independent employing the Naïve Bayes assumption. This assumption seems reasonable, as the queried value cell does not necessarily depend on the queried column header. Thirdly, we employ the Naïve Bayes assumption to consider that, b_{ch} and b_v are conditionally independent of q_{ch} and q_v respectively. Finally, the last approximation comes from the fact that an attribute-value query (q_{ch}, q_v) is just a Boolean AND query, $q_{ch} \wedge q_v$, and according to [14], AND relevance probability can be better approximated using the minimum rather than the product.

Finally, from Eq. (3.1) and (3.2):

$$P(\mathcal{R} \mid q_{ch}, q_v) \approx \max_{b_{ch}, b_v \sqsubseteq x} P(b_{ch})\, P(b_v \mid b_{ch})\, \min(P(q_{ch} \mid b_{ch}), P(q_v \mid b_v)) \tag{3.3}$$

The first factor, $P(b_{ch})$, is a prior probability for the position of the column header in the x-axis and the second, $P(b_v \mid b_{ch})$, is the conditional probability of the position and geometry of b_v, given the position and geometry of the corresponding header, b_{ch} . The last two factors, $P(q_{ch} \mid b_{ch})$ and $P(q_v \mid b_v)$ are obtained by combining the relevance probabilities of the *pseudo-words* involved in q_{ch} and q_v , directly provided by the PrIx of the table image x. The bayesian network representing this probabilistic model can be seen in Fig. 3.2.

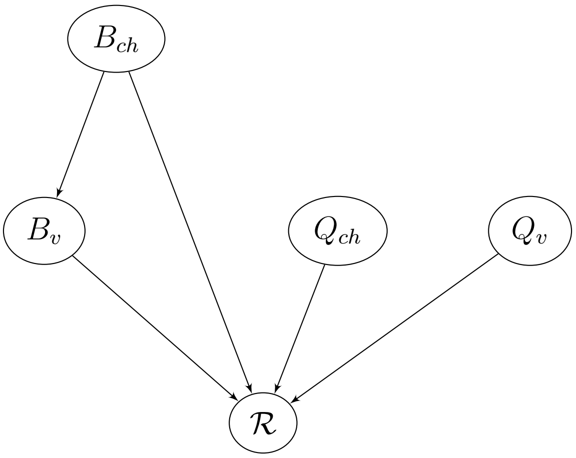


Figure 3.2: Bayesian network describing the probabilistic model to compute the relevance of a column query.

3.2 Row-header Cell-Value Queries

Now, we are going to consider the problem of performing row queries. With this purpose, we want to compute the relevance probability $P(\mathcal{R} \mid x, q)$, where x is a table image and q is a query. In this subsection we assume $q = (q_{rh}, q_v)$, where q_{rh} is a row header "attribute" query and q_v a value or cell query. For simplicity, we assume the regions of x matching q_{rh} and q_v are small bounding boxes, b_{rh}, b_v, respectively, each tightly containing the words specified by the corresponding query. Therefore:

$$P(\mathcal{R} \mid x, q) \;\equiv\; P(\mathcal{R} \mid q) \;\equiv\; P(\mathcal{R} \mid q_{rh}, q_v)$$

Then, it can be seen that Eq. (3.3) can be straightforwardly adapted to perform row queries as:

$$P(\mathcal{R} \mid q_{rh}, q_v) \approx \max_{b_{ch}, b_v \sqsubseteq x} P(b_{rh})\, P(b_v \mid b_{rh})\, \min(P(q_{rh} \mid b_{rh}), P(q_v \mid b_v)) \tag{3.4}$$

The first factor, $P(b_{rh})$, is a prior probability for the position of the row header in the y-axis and the second, $P(b_v \mid b_{rh})$, is the conditional probability of the position and geometry of b_v, given the position and geometry of the corresponding row header, b_{rh}. The last two factors, $P(q_{rh} \mid b_{rh})$ and $P(q_v \mid b_v)$ are obtained by combining the relevance probabilities of the *pseudo-words* involved in q_{rh} and q_v , directly provided by the PrIx of the table image x. The bayesian network representing this probabilistic model can be seen in Fig. 3.3.

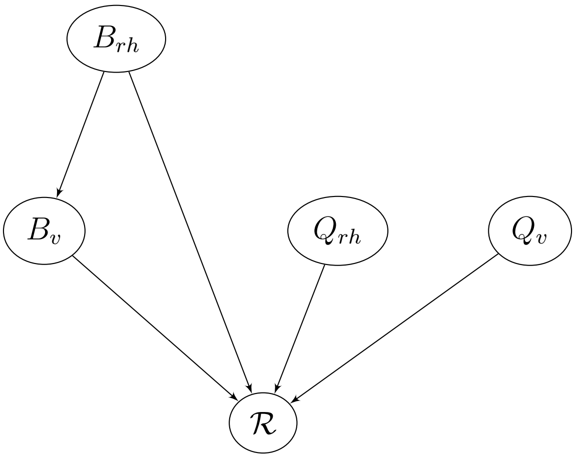


Figure 3.3: Bayesian network describing the probabilistic model to compute the relevance of a row query.

3.3 Row-column Intersection Queries

Now, we are going to consider the problem of performing row-column queries. With this purpose, we want to compute the relevance probability $P(\mathcal{R} \mid x, q)$, where x is a table image and q is a query. In this subsection we assume $q = (q_{rh}, q_{ch}, q_v)$, where q_{rh} is a row header "attribute" query, q_{ch} is a column header "attribute" query and q_v a value or cell query. For simplicity, we assume the regions of x matching q_{rh}, q_{ch} and q_v are small bounding boxes, b_{rh}, b_{ch} and b_v respectively, each tightly containing the words specified by the corresponding query. Therefore:

$$P(\mathcal{R} \mid x, q) \;\equiv\; P(\mathcal{R} \mid q) \;\equiv\; P(\mathcal{R} \mid q_{rh}, q_{ch}, q_v)$$

Following the approximation presented in [14] to combine probability distributions, it can be seen that $P(\mathcal{R} \mid q_{rh}, q_{ch}, q_v)$ can be approximated as:

$$P(\mathcal{R} \mid q_{rh}, q_{ch}, q_v) \;\approx\; \min(P(\mathcal{R} \mid q_{rh}, q_v), P(\mathcal{R} \mid q_{ch}, q_v)) \tag{3.5}$$

where $P(\mathcal{R} \mid q_{rh}, q_v)$ and $P(\mathcal{R} \mid q_{ch}, q_v)$ denote the probability associated to the row and column queries respectively of which we want to obtain the intersection.

3.4 Extension to information extraction

Despite our proposal is mainly aimed at information search, it can be naturally extended to perform information extraction with the help of wildcard queries.

A wildcard query could be defined as a query where we find a sequence of characters of any length followed by the character "*", which denotes another sequence of characters of any length. Given this information, the task of performing information extraction can be seen as a regular tabular query where the queried value-cell q_v is the symbol "*".

Please note that, when querying for $P(\mathcal{R} \mid q_{rh}, q_{ch}, q_v = *)$, the value of q_v would be expanded into all the possible *pseudo-words* that appear in the PrIx, leading this to retrieve a probability distribution of hypothesis for that cell instead of a plain transcription, which is the usual format at this task. Of course, this format is not compatible with a standard database, given that it provides a probabilistic transcription. However, it opens the need of working and developing probabilistic databases for this type of tasks, where the entries would be formed by the *pseudo-words* found in that region along with their probabilities. By doing this, the users could retrieve the n-best BBs for each cell, as is shown in Eq. (2.4), (and obtain as a by-product the *pseudo-words* attached to them), where the value of n would depend on the user needs for each application.

As a first approximation to this paradigm, we have decided to retrieve only the BB of higher probability for each cell. By doing this, we obtain a format that is directly suitable for standard databases. Firstly, we could calculate the relevance probability when extracting information from an intersection between a column query q_{ch} and a row query q_{rh} as:

$$P(\mathcal{R} \mid q_{rh}, q_{ch}, q_v = *) \equiv \max_{w \epsilon W} \min(P(\mathcal{R} \mid q_{rh}, q_v = w), P(\mathcal{R} \mid q_{ch}, q_v = w)) \quad (3.6)$$

where W denotes all the *pseudo-words* that appear in the PrIx. Finally, the best extraction for that cell can be obtained as a by-product:

$$\hat{q}_v = \arg \max_{w \epsilon W} \min(P(\mathcal{R} \mid q_{rh}, q_v = w), P(\mathcal{R} \mid q_{ch}, q_v = w)) \quad (3.7)$$

CHAPTER 4
Experiments and Results

4.1 Corpus

The employed corpus to assess the ideas proposed in this thesis is the first version of the HisClima database [11]. This corpus is compiled from the logbook of *Jeannette*, a ship which sailed the Arctic ocean from July of 1880 until February of 1881. During this expedition, the sailors recorded Climatic information several times a day: wind speed, temperature, coordinates, the form of the clouds, etc.

In this logbook, we find that each annotated date typically corresponds to two contiguous pages: the left page, which contains tabular information about the weather conditions at certain times of the day, and the right page, which contains miscellaneous information about the events that occurred during that day. Thus, this corpus is composed of 419 pages, where 208 correspond to tabular pages, and the other 211 correspond to descriptive pages. Examples of both types of pages can be seen in Fig. 4.1. Clearly, in this work we will only focus on the tabular pages.

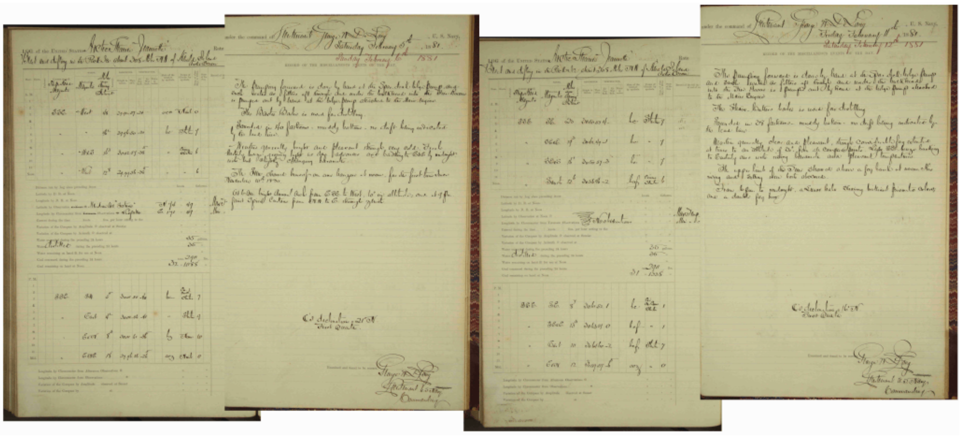
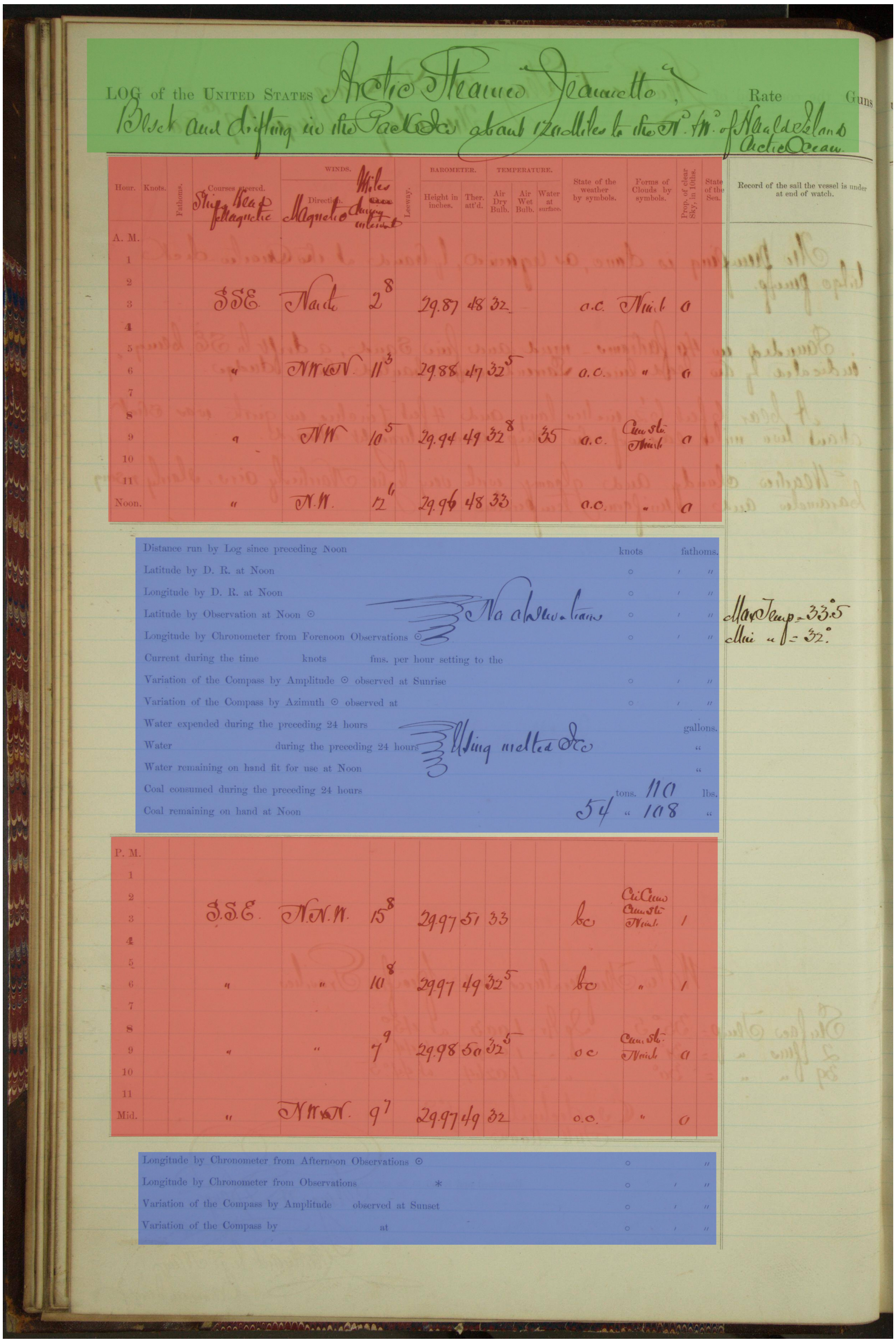


Figure 4.1: Examples of tabular and miscellaneous pages that conform the collection.

Now, if we take a look at Fig. 4.2 we can observe that each tabular page is composed of five regions: Firstly, the header of the document, where we find the date and title of the page. Secondly, we find the first table region, where we can see the column headers of the table, which correspond to the Climatic conditions that were recorded several times during a day. In this table region, the values annotated correspond to the measurements performed before noon, as it is denoted by the header “A. M.”, present in the first column of the table. In third place, we find a structured region where there are recorded other relevant events, which were measured less frequently than the information found in the table regions. In fourth place we find another table region, which shares the columns headers found in the first table and accounts for the measured events in the afternoon, as it is denoted by the header “P. M.”, present in the first column of the table. Finally, we can find another structured region. In this work, we will only focus on the table regions.



LOG of the UNITED STATES Arctic Steamer Jeannette, Rate Guns

Beset and drifting in the Pack Ice about 120 Miles to the N. W. of Herald Island Arctic Ocean.

Hour.	Knots.	Fathoms.	Courses steered. Ship Head Magnetic	WINDS. Direction. Magnetic	Miles Force Cheng. interest	Leeway.	BAROMETER. Height in inches.	BAROMETER. Ther. att'd.	TEMPERATURE. Air Dry Bulb.	TEMPERATURE. Air Wet Bulb.	TEMPERATURE. Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.	Record of the sail the vessel is under at end of watch.
A. M.																
1																
2																
3			SSE	Nact	2^{8}		29.87	48	32			o.c.	Nimb	0		
4																
5																
6			"	NW&N	11^{3}		29.88	47	32^{5}			o.c.	"	0		
7																
8																
9			"	NW	10^{5}		29.94	49	32^{8}		35	o.c.	Cum str Nimb	0		
10																
11																
Noon.			"	N.W.	12^{6}		29.96	48	33			o.c.	"	0		

Distance run by Log since preceding Noon		knots		fathoms.
Latitude by D. R. at Noon		°	′	″
Longitude by D. R. at Noon		°	′	″
Latitude by Observation at Noon ⊙	No observations	°	′	″
Longitude by Chronometer from Forenoon Observations ⊙		°	′	″
Current during the time knots fms. per hour setting to the				
Variation of the Compass by Amplitude ⊙ observed at Sunrise		°	′	″
Variation of the Compass by Azimuth ⊙ observed at		°	′	″
Water expended during the preceding 24 hours				gallons.
Water during the preceding 24 hours	Using melted Ice			“
Water remaining on hand fit for use at Noon				“
Coal consumed during the preceding 24 hours		tons.	110	lbs.
Coal remaining on hand at Noon		54 “	108	“

Max Temp = 33°.5
Min " " = 32°

Hour.	Knots.	Fathoms.	Courses steered.	Direction.	Miles	Leeway.	Height in inches.	Ther. att'd.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
P. M.															
1															
2															
3			S.S.E.	N.N.W.	15^{8}		29.97	51	33			bc	Ci Cum Cum str Nimb	1	
4															
5															
6			"	"	10^{8}		29.97	49	32^{5}			bc	"	1	
7															
8															
9			"	"	7^{9}		29.98	50	32^{5}			oc	Cum str Nimb	0	
10															
11															
Mid.			"	NW&N.	9^{7}		29.97	49	32			o.c.	"	0	

Longitude by Chronometer from Afternoon Observations ⊙	°		″
Longitude by Chronometer from Observations *	°	′	″
Variation of the Compass by Amplitude observed at Sunset	°	′	″
Variation of the Compass by at	°	′	″

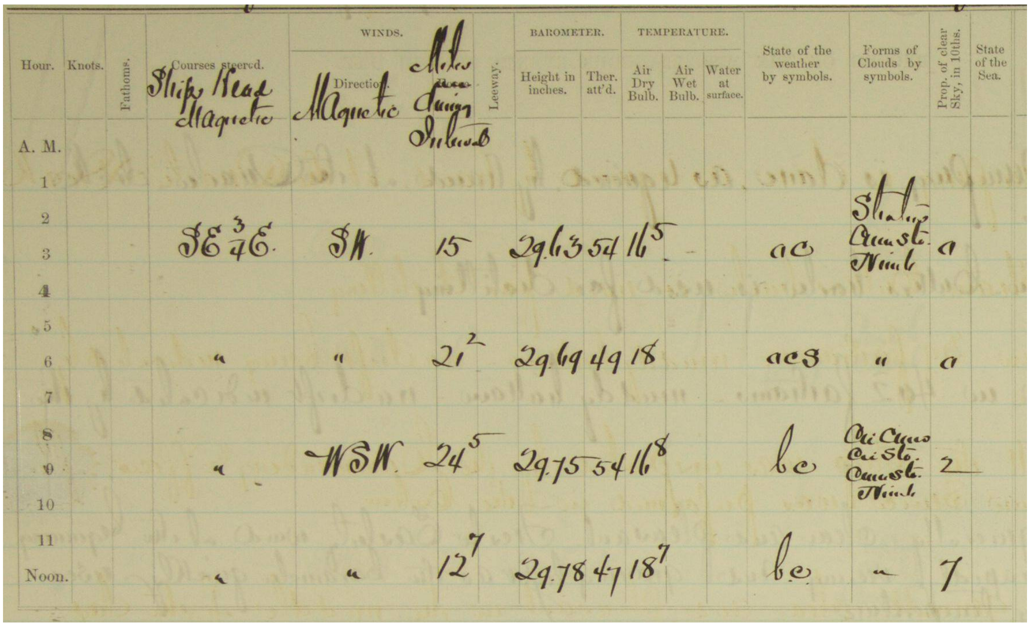
Figure 4.2: Example of tabular page from the *Jeannette* logbook. The title region is represented in green, the table regions are coloured in red and the register regions are denoted in blue.

Now, if we take a closer look at Fig. 4.3, we can observe some of the challenges that should be taken into account when working with the tables of this collection.

Firstly, the width of each column is completely different. For instance, the wind direction column is wider than all the columns that account for the barometer information.

Secondly, it can be seen that frequently, when the measured value of a cell is the same as the value of the last annotated predecessor in the column, the writer of the document decided to write the quotation marks symbol (denoting that the content is the same as the content of the predecessor cell on the column) instead of writing the value.

Finally, if we observe the column whose header value is "Forms of Clouds by symbols", it can be seen that when the content of a value cell is larger than the cell size, it is split into multiple adjacent cells in the same column. It is worth noting that, when performing information search and/or extraction, we will aim to retrieve all the contents that are semantically related to the queried row, even when the content of a cell is divided into multiple cells.



				WINDS.			BAROMETER.		TEMPERATURE.						
Hour.	Knots.	Fathoms.	Courses steered. Ship's Head Magnetic	Direction. Magnetic	Miles Force during Interval	Leeway.	Height in inches.	Ther. att'd.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.
A. M.															
1															
2													Stratus		
3			SE $\frac{3}{4}$ E.	SW.	15		29.63	54	16^{5}			ac	Cumst.	a	
4													Nimb		
5															
6			"	"	21^{2}		29.69	49	18			acs	"	a	
7															
8													Ci Cumu		
9			"	WSW.	24^{5}		29.75	54	16^{8}			bc	Ci Sto.	2	
10													Cumst. Nimb		
11															
Noon.			"	"	12^{7}		29.78	47	18^{7}			bc	"	7	

Figure 4.3: Example of table region of the HisClima dataset. Some of the challenges of this collection can be found in this image. For instance, in the cell that is the intersection of the column "Courses steered" and the row "6", we can see that its value is the quotation marks. Moreover, in this table we find that the contents of two cells of the column "Forms of clouds by symbols" are split into three cells each.

4.2 Exploratory Analysis of the data

In order to decide which attributes of the data are interesting to take into account when modelling the probabilistic framework presented in Chap. 3, we have performed an exploratory analysis of the data.

Firstly, we have looked at the attributes that could be interesting to model $\mathrm{P}(b_{ch})$. The two attributes considered are the position in the x-axis and the position in the y-axis.

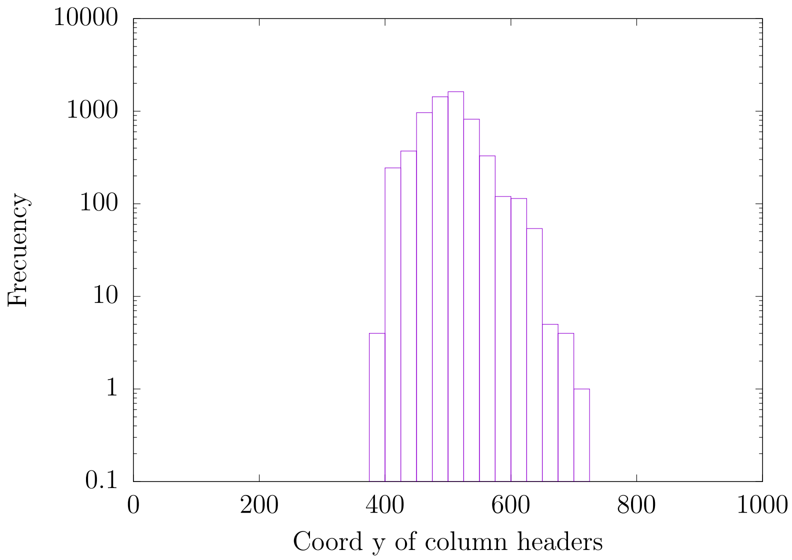


Figure 4.4: Distribution of column header BB in function of their y coordinate

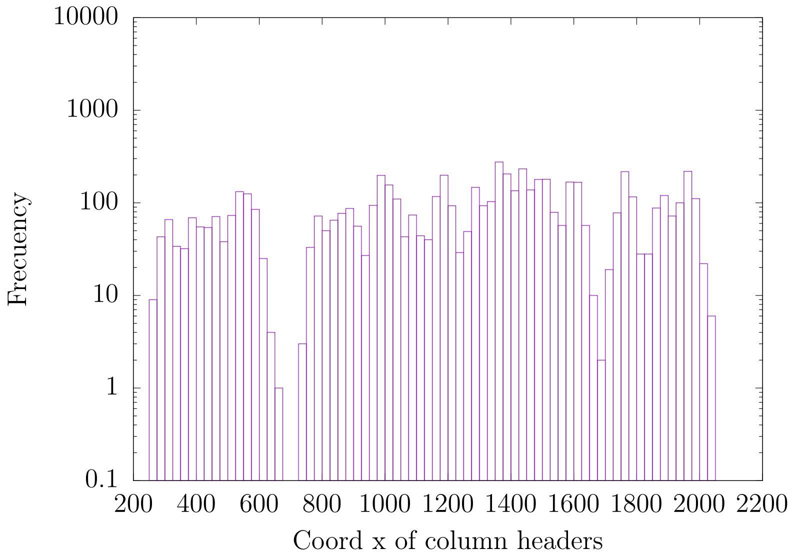


Figure 4.5: Distribution of column header BB in function of their x coordinate

If we observe the plots, it seems reasonable to discard the x-axis, as the data is normally distributed across it. Therefore, $P(b_{ch})$ could be modelled employing a univariate Gaussian distribution over the y-axis.

Secondly, as attributes that could be interesting to model $P(b_v \mid b_{ch})$ we have considered the width of the value-cell BB (denoted as w_v) and the geometric centre of it in the x-axis (denoted as cx_v). In order to generalise to all the different column shapes, we have employed the deviations of each attribute with respect to the width and geometric centre of the associated b_{ch} (w_{ch} and cx_{ch} respectively).

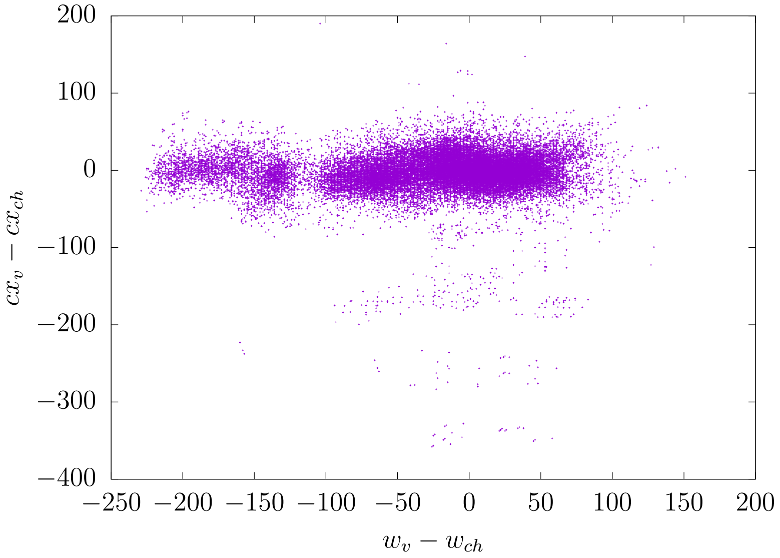


Figure 4.6: Scatter plot showing the relation between $w_v - w_{ch}$ and $cx_v - cx_{ch}$

If we observe Fig. 4.6, two differentiated clouds of points can be seen. On the one hand, if we take a look at the cloud of points that is found on the left, it can be seen that the width of the cell-values is much smaller than the width of the queried column header. This is due to the fact that the quotation marks are tiny, meanwhile the column headers might be much larger. On the other hand, if we look at the cloud of points found on the right, it can be seen that the difference between the widths is in a reasonable range and that typically, the geometric centre of the cell-values is similar to the geometric centre of the associated column headers. Taking all this information into account, it seems reasonable to model $P(b_v \mid b_{ch})$ as a two-dimensional Gaussian with diagonal covariance matrix, where one dimension would be $w_v - w_{ch}$ and the other one would be $cx_v - cx_{ch}$.

Now, we are going to take a look at the attributes that could be interesting to model $P(b_{rh})$. The two attributes that we have considered at first glance are the x coordinate and the y coordinate of b_{rh}.

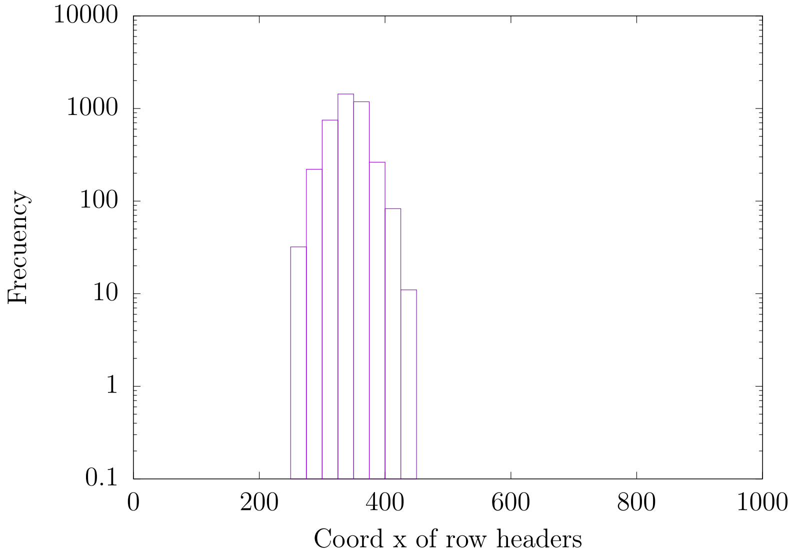


Figure 4.7: Distribution of row header BBs in function of their x coordinate

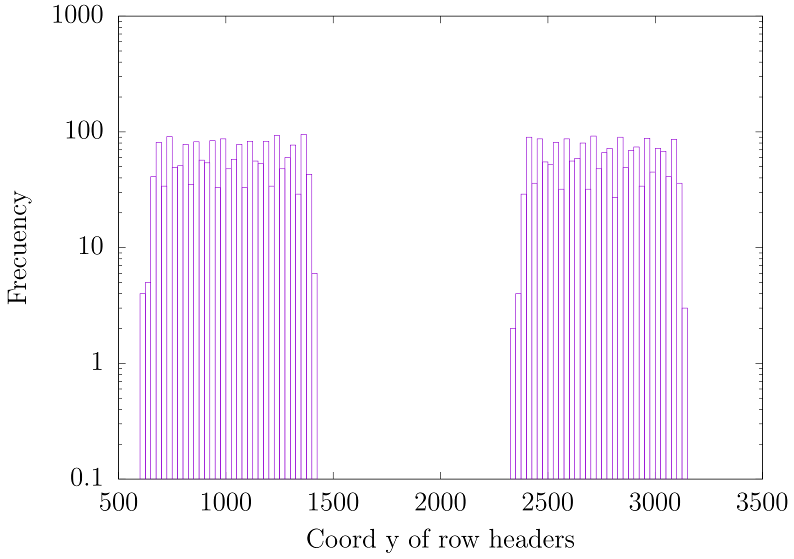


Figure 4.8: Distribution of row header BBs in function of their y coordinate

First of all, we would like to remark that the most informative dimension is the x-axis, as in this collection the row header is the first cell of each row of the table. Finally, we have decided to discard the information of the y-axis, since the data is normally distributed over the two tables that conform each page. Given this, we have decided to model $P(b_{rh})$ as a univariate Gaussian over the x-axis dimension.

Finally, let's see how could we model $P(b_v \mid b_{rh})$. Analogously to $P(b_v \mid b_{ch})$, this probability can be modelled employing $h_v - h_{rh}$ and $cy_v - cy_{rh}$, where $cy_v - cy_{rh}$ denotes the deviation between the geometric centre in the y-axis of b_v and the associated geometric centre of b_{rh}, and $h_v - h_{rh}$ represents the difference between the height of b_v and the height of b_{rh}.

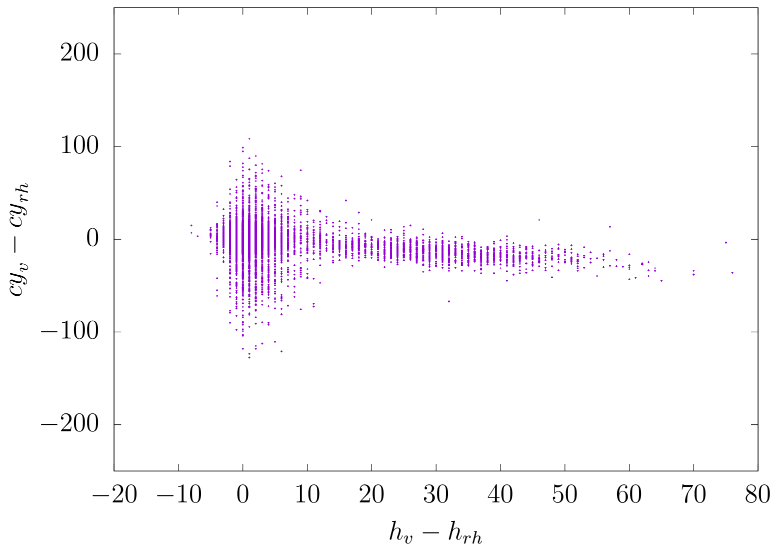


Figure 4.9: Scatter plot showing the relation between $h_v - h_{ch}$ and $cy_v - cy_{ch}$

If we observe Fig. 4.9, it can be seen that the cell values are typically centred in the y-axis with respect to cy_{ch}. However, it is not always the case, as sometimes the deviation is larger than 100 pixels. This is due to the fact that when the data does not fit into a single cell, the writer of the document decided to distribute it across the adjacent cells in the vertical axis, leading to these large deviations between cy_v and cy_{ch}. Given this information and the shape of the scatter plot, it seem reasonable to model $P(b_v \mid b_{rh})$ as a two dimensional Gaussian with diagonal covariance matrix, where the two dimensions are $h_v - h_{ch}$ and $cy_v - cy_{ch}$.

4.3 Experimental setup

To evaluate empirically the ideas proposed in this thesis, we have performed different experiments. In this section, we describe the considered partitions of the corpus, the software that has been employed to estimate the probability distributions, the performed experiments, the evaluation protocol and the input and output of our system.

First of all, from the 208 tabular pages that can be found in the HisClima database, we have made two different partitions: the train partition, which has been used to train the statistical models, and the test partition, which has been employed to assess the performance of our system. Statistics about these partitions can be found in Tab. 4.1. Please note that the row "Rel. information" accounts for the number of cell values found in each partition.

	Train	Test	Total
Pages	158	50	208
Lines	25 901	7 838	33 739
Rel. information	11 938	3 533	15 741

Table 4.1: Basic statistics of the HisClima partitions.

Secondly, as library to estimate the different probability distributions and their parameters, we have employed the Armadillo [13] toolkit.

Thirdly, to assess the ideas presented in this thesis we have performed two experiments: one aimed at information extraction and another aimed at information retrieval.

The first task that we have considered is information extraction, as it is one of the main goals in this collection. In this task, the objective is to obtain a transcription for each cell of the table, in order to fill this information into a standard database. The second experiment that we have considered is aimed at information retrieval. Concretely, we have decided to emulate the behaviour of the users when searching information at column level. For instance, one example of query would be: "In which days the direction of the wind was North East?".

As query set for the information extraction experiment, we have employed all the possible combinations between row and column headers, as we are interested in retrieving the contents of all the cells that conform each table, meanwhile in the information retrieval experiment we have employed as query set all the existing tuples <column, keyword> present in the GT.

Moreover, given the different objectives of the two experiments, we would like to remark that we have employed different criteria to evaluate them.

On the one hand, the evaluation of the information extraction experiment is performed at cell level. In this experiment, a match is considered a *true positive* (TP) when the retrieved content of the cell by the system is exactly the same as the content of the cell found in the *ground truth* (GT). Otherwise, the match is considered a *false positive* (FP). Finally, if the content of a cell in the GT is not retrieved by the system, this is considered as a *false negative* (FN).

On the other hand, the evaluation of the information search experiment is performed at column level. Given this, a match is considered a *true positive* (TP) when the retrieved keyword by the system appears in the same column in the GT. Otherwise, the match is considered a *false positive* (FP). Finally, if a keyword in a column in the GT is not retrieved by the system, this is considered as a *false negative* (FN).

Now, we are going to discuss the input of our system. Unfortunately, at the moment of writing this thesis we do not have reliable PrIx for this collection. However, we have a 1-best transcription, which accounts for the best transcription of our system for each line. Nevertheless, given that in this collection the lines are annotated at cell level and taking into account that each line typically holds only one *pseudo-word*, we can replace the use of the PrIx by the 1-best transcription in this particular case. Given this, the probabilities of the probabilistic framework are estimated employing the BBs associated to the line where appears the queried *pseudo-words*. Moreover, the use of the 1-best transcription makes our results comparable to the ones presented in [11], which we are going to consider as baseline in our experiments.

Finally, as output for our system, in the case of information extraction we are going to return a hypothesis for each cell of each table, along with its relevance probability, meanwhile for the information retrieval experiment we are going to return a list of pages where the queries match the system hypothesis, along with a relevance probability for each page.

4.4 Information extraction

In order to perform information extraction at cell level, we have followed these steps:

For each value-cell of the table, firstly we have retrieved the BB associated to the line of the *pseudo-word* that matches q_{ch}. This fact is sound in this collection because we know that there is at least one unique and distinctive *pseudo-word* for each column. Secondly, we have performed the analogous procedure to locate the row header BB associated to q_{rh}. Thirdly, we have retrieved all the BBs present on the page and we have considered them as possible b_v. Then, we have calculated $P(\mathcal{R} \mid q_{rh}, q_{ch}, q_v)$ as detailed in Chap. 3 for each b_v. Next, we have approximated the distribution that accounted for all the possible q_v in that precise cell by the q_v which leads to the highest $P(\mathcal{R} \mid q_{rh}, q_{ch}, q_v)$ employing Eq. (3.6). Finally, for those cells whose BB of maximum probability is a quotation mark we have performed a post-process to substitute them by their semantic value. With this purpose, we would consult the q_v associated to the maximum RP hypothesis of the precedent cell in the vertical axis. If it is empty or it is another a quotation mark, then we would look into the next precedent cell. Otherwise, we would return as value the q_v of that cell. Moreover, we would like to remark that the final probability of the quotation mark cell is the minimum between the probabilities of all the cells which have been visited until achieving the first value cell which is not a quotation mark.

It is worth noting that, in the third step of this task, many possible configurations of Gaussian distributions could be fitted to model the different probabilities that account for $P(\mathcal{R} \mid q_{rh}, q_{ch}, q_v)$. In order to determine the best configuration for this collection, we have performed different experiments. Results can be found in Tab. 4.2.

Approach	P	R	F$_1$
Baseline	0.79	0.79	0.79
1G col-value and 1G row-value v1	0.68	0.49	0.57
1G col-value and 1G row-value v2	0.70	0.51	0.59
11G col-value and 24G row-value	0.74	0.77	0.76
11G col-value and 1G row-value	**0.81**	**0.80**	**0.81**

Table 4.2: Information extraction precision, recall and F$_1$ at cell level.

For all the experiments, we have employed a univariate Gaussian to model $P(b_{ch})$ and another to model $P(b_{rh})$. Now, we are going to detail the different configurations for $P(b_v \mid b_{ch})$ and $P(b_v \mid b_{rh})$ that have been tried. Firstly, we have modelled $P(b_v \mid b_{ch})$ employing a Gaussian and $P(b_v \mid b_{rh})$ employing another Gaussian, as was proposed in Sec. 4.2.

In this first experiment, we tried two different training data to estimate $P(b_v \mid b_{rh})$. In the first version, denoted as “v1”, we employed all the data that is shown on Fig. 4.9 to estimate the parameters of the Gaussian, meanwhile for the second version, denoted as “v2”, we only employed the data which was not present in the column of “Form of clouds”. The main idea behind the second version is that some lines in that column are annotated as the row where they semantically belong, which is different from the physical row where the lines are found. This is due to the fact that, when the contents they wanted to write in that cell were larger than it, they decided to split them into the vertical adjacent cells. If we observe Tab. 4.2, it can be seen that we have achieved better results with the second configuration. Given this, in the following experiments we will employ the second criteria to choose the training data to estimate $P(b_v \mid b_{rh})$.

Nevertheless, we would like to remark that the results were very poor in both versions, being the F_1 score 22 and 20 points worse than the baseline results. In view of this, we decided to observe the system's predictions and found that, frequently, our system returned as best hypothesis the adjacent cell (either in the vertical or horizontal axis). This was due to the fact that trying to model $P(b_v \mid b_{ch})$ and $P(b_v \mid b_{rh})$ employing a unique Gaussian for each was too demanding for the system (for instance, a large deviation of a value cell with respect to its column header in a wide column might be negligible, meanwhile the same deviation in a smaller column could be critical).

Then, we tried the opposite experiment, employing a different Gaussian for each column (11 Gaussians in total to model $P(b_v \mid b_{ch})$) and a different Gaussian for each row (24 Gaussians in total to model $P(b_v \mid b_{rh})$). It can be seen that we improved 17 points the F_1 score with respect to the first approximation. Again, we took a closer look at the mistakes that were made by our system and we found that, in the rows which were not typically filled in the training set, the Gaussian parameters were estimated with fewer data and tended to overfit.

As a possible solution to this problem, we repeated the experiment but employing a unique Gaussian to model all the rows. We thought that this assumption was sound, given that we have modelled $P(b_v \mid b_{rh})$ employing relative deviations and that all the rows have the same height and length. It can be seen that with this approach we have achieved our best performance (81 of F_1 score).

Now, let's take a look at where is the system failing. We find two main problems: the skew and the multi-row contents of the column associated to the header "form of clouds".

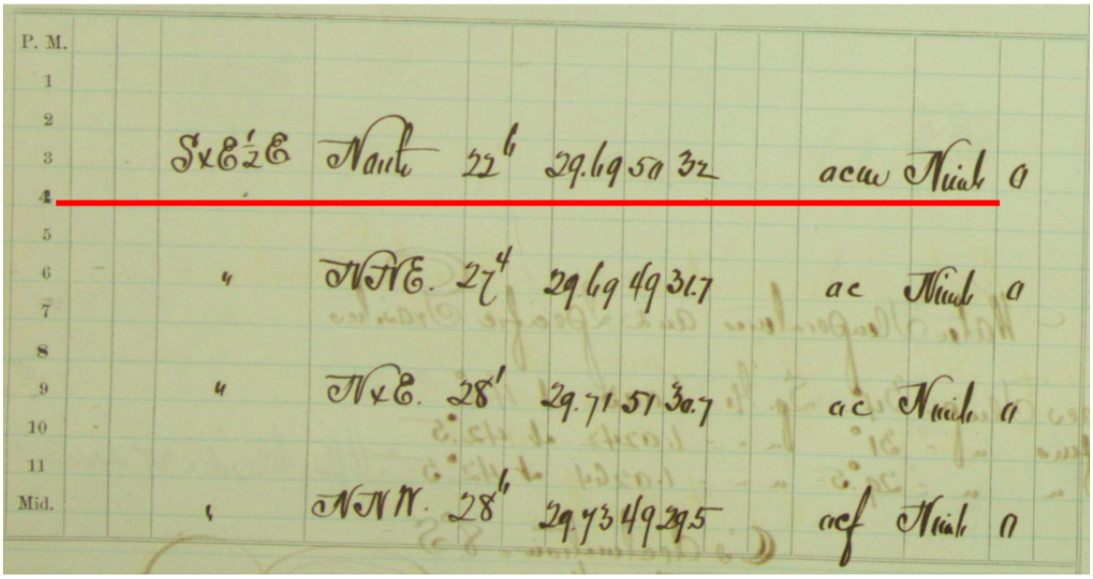


Figure 4.10: Example of HisClima page with severe skew. The red line is parallel to the x-axis.

Firstly, there are some pages that present skew, which makes difficult for our system to make the correct prediction. For instance, if we take a look at Fig. 4.10, it can be seen that if we draw a horizontal line at the position of the fourth row, it seems more reasonable to assume that the cell value "Nimb" is accounting for the row header "4" instead of "3", which is its real row header. Despite there exist techniques to alleviate this problem [6] which could work on an homogeneous corpus like the HisClima database, this problem could not be easily solved for a more heterogeneous corpus where the structure of the tables is handwritten and erratic, such as the Passau Tables collection.

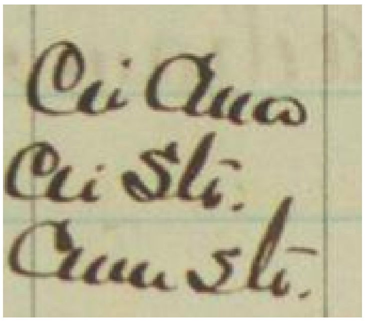


Figure 4.11: Example of multi-row cell contents.

Secondly, the other major problem is the cells which contents are distributed over the adjacent vertical cells. For instance, in Fig. 4.11, we find that the content that semantically belongs to the central cell is distributed into three cells. Our system, would retrieve each line in each physical cell, accounting for three *false positives*, as the first and last cell are annotated as empty in the GT and the middle cell is annotated as the combination of the three cells, and a *false negative*, as we have not retrieved the cell contents. This fact hinders the performance of our system severally. As a possible solution to this problem, we could combine the contents of the upper and lower cell into the middle cell in a post-process. However, this task is also challenging, as not all the contents of cells in that column are split into multi-rows.

4.5 Information search

The process to perform information retrieval at column level is the following:

Firstly we have retrieved the BB associated to the line of the *pseudo-word* that matches q_{ch}. Secondly, we have retrieved all the BBs present in the page that correspond to q_v. Finally, we have calculated $P(\mathcal{R} \mid q_{ch}, q_v)$ as detailed in Chap. 3 for each b_v. Results can be found in Tab. 4.3.

It can be seen that our system improves the baseline AP by 3 points and the mAP by 4 points. This might be due to the fact that the baseline approach only retrieves a binary result for each query (either the query is fulfilled or not by a page), meanwhile our system returns a probability for each possible query, allowing the user to choose the threshold that suites most its needs. This can be seen in Fig. 4.12, where the baseline is denoted by a perfect rectangle whereas our approach is denoted by a curve.

Approach	AP	mAP
Baseline	0.86	0.80
1G col-value and 1G row-value	**0.89**	**0.84**
11G col-value and 1G row-value	0.89	0.84

Table 4.3: Information retrieval AP and mAP at column level

Moreover, we would like to remark that our system achieves the same results employing our first approximation to the problem or the most complex approximation that yielded the best results at information extraction. This is due to the fact that, during the information extraction experiments, when employing the first approximation sometimes the best result was an adjacent cell and the system was retrieving as result the q_v with the highest probability, meanwhile the correct transcription according to the GT was usually the second or third best hypothesis. This situation motivated the use of more specific Gaussians for each column in the information extraction task. Nevertheless, as in information retrieval the user specifies the queried *pseudo-word* q_v, the system does retrieve the hypothesis that matches the performed query, and therefore, we do not face the maximisation problem that motivated the use of specific Gaussians for each column.

Finally, we would like to remark that some of the major concerns that appeared in information extraction are not a problem anymore in information retrieval. For instance, as we are evaluating this task at column level, the skew problem is alleviated given that we do not take rows into account, and therefore, row skew does not hinder the performance of our system. Furthermore, the multi-row cells are no longer a problem, given that these rows belong to the same column and therefore, they will be retrieved correctly.

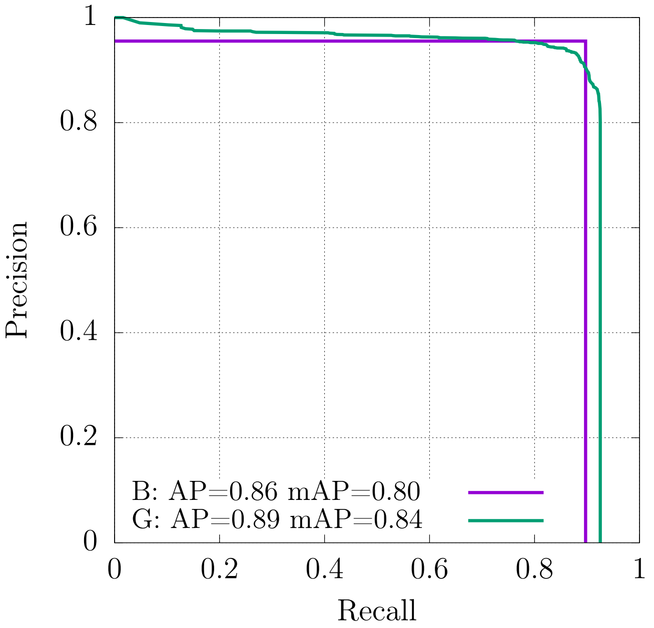


Figure 4.12: RP curves for the information retrieval task. “B” denotes the baseline method, meanwhile “G” denotes our first approach, based in employing a Gaussian to model $P(b_v \mid b_{ch})$ and another Gaussian to model $P(b_v \mid b_{rh})$.

CHAPTER 5
Conclusions and future works

5.1 Conclusions

At the beginning of this thesis, we had proposed three objectives to accomplish: Proposing a probabilistic framework for tabular queries, trying different machine learning models to estimate the required probabilities and measuring the performance of the proposed systems when searching and extracting information.

For the first objective, we have defined a probabilistic layout agnostic framework that allows performing tabular column queries, row queries and the intersection of both. Moreover, we have discussed how to adapt it to perform information extraction.

For the second objective, we have tried different configurations of Gaussian distributions to model $P(b_v \mid b_{ch})$ and $P(b_v \mid b_{rh})$ at information extraction and information retrieval tasks. We have seen empirically that the use of multiple Gaussians accounting for the different columns of the tables yields a huge improvement with respect to employing a unique Gaussian to model $P(b_v \mid b_{ch})$ at the information extraction task. However, this gain is only present at information extraction, as we have achieved the same performance when employing a unique Gaussian for $P(b_v \mid b_{ch})$ at the information retrieval experiment.

For the third objective, we have measured the performance of the system over these tasks and we have discussed which are the main advantages of our system when performing information search and extraction compared to the baseline, but also which are the main concerns that the system is facing when performing those tasks.

Finally, the achievement of these three minor objectives, as well as the favourable results in information search and extraction, makes possible to affirm that we have achieved the main objective of this work: proposing a probabilistic framework that substitutes the heuristics employed in the baseline method for well known statistical models, achieving comparable performance.

5.2 Future works

As possible future works that could be performed we find:

- Assess the ideas presented in this thesis over a more heterogeneous corpus such as the Passau Tables collection and compare our results against the proposed baseline [7].
- Try other machine learning models to estimate the different probabilities of the probabilistic framework. In this work, we have focused on the use of Gaussian distributions due to their model explainability and their easiness of estimation and use. However, other machine learning models could have been used for this task, as for example, a Neural Network such as the Perceptron [12], a Gradient Boosting Regressor [5] or an SVM [3].
- Perform information search and extraction over the non-table structured regions of the pages. In this work, we have only aimed at searching and extracting information from the handwritten tables. However, there are other structured regions, below the tables, that also contain information that might be relevant for the users.
- When available, repeat the experiments of this work employing *probabilistic indices*. Despite being sound the use of the 1-best transcription instead of a probabilistic index for this collection, we hope achieving better results with the PrIx, as where the 1-best fails to transcript the contents of the image there might be other accurate hypotheses for that region in the PrIx.
- Assess the performance of the system when performing information retrieval at cell level. For instance, instead of asking for a concrete temperature on a concrete day, ask for a concrete temperature at a determined moment of the day.
- Develop range queries in order to allow the users to ask for ranges of values instead of a unique keyword. This would allow the users to ask for a range of temperatures instead of asking for a concrete temperature. This type of query would answer questions like “In which days the air temperature was between 20 and 30 degrees?”.
- Develop a post-process to mitigate the effects of the multi-row cell contents for information extraction.

Bibliography

[1] Christopher M Bishop. *Pattern recognition.* Vol. 128. 9. 2006.

[2] Stéphane Clinchant et al. “Comparing machine learning approaches for table recognition in historical register books”. In: *2018 13th IAPR International Workshop on Document Analysis Systems (DAS).* IEEE. 2018, pp. 133–138.

[3] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20.3 (1995), pp. 273–297.

[4] Leo Egghe. “The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations”. In: *Information Processing & Management* 44.2 (2008), pp. 856–876.

[5] Jerome H Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232.

[6] Basilios Gatos, Nikos Papamarkos, and Christodoulos Chamzas. “Skew detection and text line position determination in digitized documents”. In: *Pattern Recognition* 30.9 (1997), pp. 1505–1519.

[7] Eva Lang et al. “Probabilistic indexing and search for information extraction on handwritten german parish records”. In: *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR).* IEEE. 2018, pp. 44–49.

[8] Thibauld Nion et al. “Handwritten information extraction from historical census documents”. In: *2013 12th International Conference on Document Analysis and Recognition.* IEEE. 2013, pp. 822–826.

[9] Joan Puigcerver. “A probabilistic formulation of keyword spotting”. PhD thesis. Universitat Politècnica de València, 2018.

[10] Stephen Robertson. “A new interpretation of average precision”. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* 2008, pp. 689–690.

[11] Verónica Romero and Joan Andreu Sánchez. “The HisClima database: historical weather logs for automatic transcription and information extraction”. In: *2020 25th International Conference on Pattern Recognition (ICPR).* IEEE. 2021, pp. 10141–10148.

[12] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386.

[13] Conrad Sanderson and Ryan Curtin. “Armadillo: a template-based C++ library for linear algebra”. In: *Journal of Open Source Software* 1.2 (2016), p. 26.

[14] Alejandro H Toselli et al. “Probabilistic multi-word spotting in handwritten text images”. In: *Pattern Analysis and Applications* 22.1 (2019), pp. 23–32.

[15] Alejandro Héctor Toselli et al. “HMM word graph based keyword spotting in handwritten document images”. In: *Information Sciences* 370 (2016), pp. 497–518.

[16] Enrique Vidal, Alejandro H Toselli, and Joan Puigcerver. “A probabilistic framework for lexicon-based keyword spotting in handwritten text images”. In: *arXiv preprint arXiv:2104.04556* (2021).