Specializations for the Peruvian Professional in Statistics: A Text Mining Approach

Luis Cajachahua Espinoza
UNI, Perú
lcajachahua@gmail.com

Andrea Ruiz Guerrero
UC, Colombia
randreag@gmail.com

Tomás Nieto Agudo
UCLM, España
Tomas.nieto.agudo@gmail.com

Abstract

The objective of this study was to identify the specialization profiles which are most required by companies and organizations in Lima, through the analysis of job postings published in the Internet. Text Mining techniques were used to extract relevant information and to identify some generic skills for the Peruvian statisticians.

For purposes of this study, we analyzed 2,809 job postings published in the Blog "Estadísticos de Perú" [2], between 2009 and 2014. We have identified many requirements, knowledge and specific skills that companies and organizations were looking for. After that, job postings were segmented using Singular Value Decomposition (SVD) of the Terms and Documents Matrix. In addition, five segments were discovered, corresponding to specific competency profiles of statisticians, where each one has different types of knowledge and specific skills.

Keywords: Job postings, Statistician, Professional, Competencies, Abilities. SVD, Clustering, Text Mining.

1 Introduction

The employment trends are changing a lot in recent years. A report published by the social network LinkedIn in 2014, after analyzing 259 million professional profiles, have identified ten professions that did not exist five years ago, but they are very popular today [11, 10]. This produces great uncertainty about the future of young people job opportunities.

On the other hand, there are many careers having accelerated growth in recent years. One of those careers is Statistics. According to reports in several countries around the world, the annual demand for professionals in Statistics has been increasing until having the highest employment rate. One example is Spain, where Statistics is the second career with the lowest unemployment rate in the country [6].

Statisticians are also required in Brazil [8], United States [1] and many other countries. According to another report, made by LinkedIn, statistical skills and data analysis are at the top of the 25 skills most sought by companies in the majority of countries considered in the study [9].

Considering these facts, there are some very interesting questions: What kind of statistics professionals are seeking companies and organizations? Have these requirements changed in recent years? Is there a unique statistician profile, or are several types? Where can we find useful information to clarify these doubts? We tried to answer these questions through analysis of job postings.

2 Background

To understand the demand for professionals and the skills required, we need to find some useful information sources. Previous research related to the issue, were made through in-depth studies, talking with some subject experts [14].

On the other hand, a group of Italian students developed a segmentation technique based on centroids [4] on the database of jobs for college SOUL (University Orientation and Job System, a network that contains jobs posted by 8 different universities in Italy) where they took more than 1,650 job postings. All kinds of them were analyzed, resulting segments from all university careers.

Another related work is the iSchool of Illinois, where they performed a segmentation analysis of Indeed job postings, in order to find the profiles that are most demanded for their students in these subjects [15]. In this case, 15,000 job postings were analyzed, all of them related to professionals in the data analysis field. But, segmentation was performed inside the contents of each job posting, so the resulting segments are referred to generic skills for all professionals.

The two last studies aimed not only to identify the most requested profiles, but also see the status of the current job market and its evolution over time, finding important patterns and can be implemented as actions either within the company or college.

2.1 Objectives

The main objectives of this study are:

- Identify the more important requirements, competencies and demands that companies include in their job postings.
- Detect the existence of professional profiles through all the job postings available through text mining techniques.
- Compare the evolution of the requirements and skills by dividing the dataset in two periods (2009-2011 and 2012-2014).

Once all previous goals achieved, we can make some recommendations to the agents involved in the job market: companies, educational institutions and potential employees, statisticians.

2.2 Limitations

By the nature of the study, it should be noted limitations implied in its realization:

- The main information source is the Blog where the job postings are published. If there were errors or omissions in the posts, they will influence the accuracy of the results.
- There are job opportunities that are not being published, causing a bias in the analysis results. Moreover, many leadership and senior positions are sent to headhunting companies. Consequently, they could not be included in this analysis.
- The postings are mostly from companies and organizations located in the city of Lima. Peru is still a very centralized country, nearly a third of Peruvian population lives in Lima, so the results could not be extrapolated to the whole country.

3 Methodology

According to the literature reviewed, there are several methods of text analysis, but these methods work well in other languages, so we needed to adapt some tools to Spanish. On the other hand, our aim, unlike previous studies, is to segment the job postings, in order to know the different types of specialties for a statistician.

3.1 Study scope

The population considered was formed by 2,809 job postings published in the blog "Estadísticos de Perú" [2]. All the postings were analyzed, so it was unnecessary to use sampling techniques. The number of postings published per year is shown in the next graph.

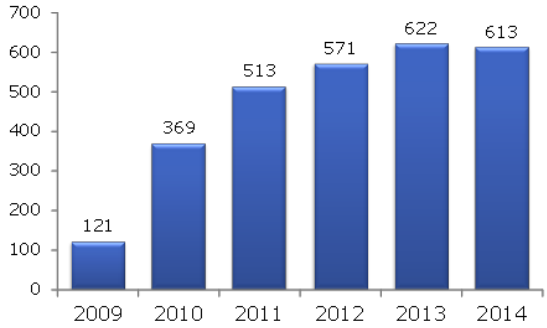


Fig. 1. Job postings per year

3.2 Text Mining

As a part of Data Mining, Text Mining is the intensive process of information extraction, where a user interacts with a collection of documents using specialized analysis tools. As a process, it deals with the discovery of knowledge in the content of several texts and after passing through several stages.

Text Mining seeks to extract useful information from multiple data sources through the identification and exploration of interesting patterns. One remarkable difference with numeric data analysis is that the documents analyzed do not have a defined structure. That is why in text mining the pre-processing tasks are very important. These operations are focused on the features identification and extraction of natural language and are responsible for transforming unstructured data in a structured intermediate format.

Text mining is used for:

- **Classify and organize documents based on their content:** With the information overload in companies, it is necessary a

method to facilitate the classification of documents that enter daily to the system. Text mining has several algorithms to do this automatically using index classification.

- **Organize depots for search and retrieval:** This problem spots the need of an efficient system search, through the submission of a request for recovering specific information. This query sends keywords to help identify the documents that best fit, sorts by relevance and the best matches are displayed. There are techniques that help to measure the similarity between documents in order to calculate the similarities and return information.
- **Automated addition and comparison of information:** Many times, when researchers have many documents on the same subject, it is necessary to group the information automatically to facilitate analysis. Text clustering is a useful technique to build the groups in these cases.
- **Extract relevant information from a document:** Text mining has methods that deals with unstructured texts, analyzes them and identifies groups of concepts. That is, it transforms plain texts into valuable and relevant knowledge.
- **Prediction and evaluation:** One of the concerns expressed sophisticated text mining is to create predictive models and evaluation from textual information that you count. These models are based on a model already raised issues of modeling and assembly, to predict for new documents entering the collection items or more suitable groups according to their contents. This type of problem is one of the most common text mining.

3.3 Text Mining Elements

Text Mining, as many other disciplines, have some recognizable elements that characterize it.

- **Repository of documents:** Any set of documents containing text, regardless of size, can be 10 or 100 billion texts. One of the main sources of documents, with more than 12 million items open to the public, with a wide variety of subjects and in different languages is PubMed. These characteristics have become one of the databases most used by computer professionals in data analysts or interested in the implementation of text mining tasks on a large scale. This collection is dynamic and are added over 40,000 items biomedical each month [17]. In a collection of this size, try to correlate the data between documents, mapping relationships or identify trends, could be extremely complex and demanding, in terms of time and machine. But there are some techniques that perform these tasks automatically that improve the speed and efficiency in the analysis.
- **Document:** For practical purposes, a document is a unit of text data (e.g. news, a report of business, emails, research articles, manuscript, stories, tweets, books, among others).
- **Corpus:** A collection of documents, usually stored electronically and on which the analysis is performed. Its elements are known as documents which store the current text and the local metadata.
- **Terms and documents matrix:** It is the most common way to represent text for future comparisons. This matrix is composed of document ID's as rows and terms as columns. Its elements are the frequencies of each term within that document.
- **Vector space model:** It is a matrix whose coefficients are functions of term frequency.

3.4 Text Mining Tools

On this study, we used R libraries and SAS Text Miner in order to obtain the results, because each one offers some advantages and useful tasks that the other one doesn't have. Another reason to choose these platforms is that the other ones do not have text Stemming and Lemmatization tools in Spanish. We can see a comparison of these tools in the next diagram:

Tasks	R*	SAS Text Miner
Ortographic Correction	✔	✔
Text Filtering	✔	✔
Multi-Words		✔
Lemmatization	✔	✔
Stemming	✔	
Term Matrix	✔	
SVD Decomposition	✔	✔
Text Segmentation		✔
Word Clouds	✔	

Fig. 2. Comparison of R and SAS Text Miner Tasks

Following this comparison, we decided to use both packages. R to clean the data and generate Word clouds for the segments and SAS Text Miner to the SVD decomposition and Segmentation.

The scheme of the Text Mining process is shown in the following image:

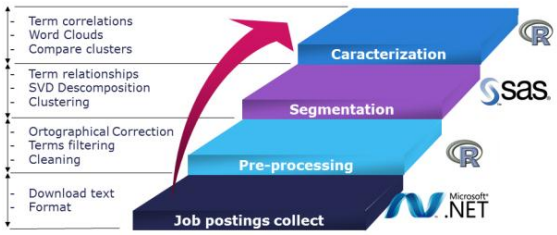


Fig. 3. Tasks and tools used in the analysis

In the terms filtering step, some stopwords were used, in order to avoid some obvious findings, like statistics, statistician, job, salary, enterprise, etc. ("estadística", "estadístico", "empleo", "salario", "empresa", etc.) Then, we performed the SVD decomposition and finally, the text clustering step. After this process, we obtained some interesting findings, which are explained in the next section.

4 Results

After textual analysis, we can answer the research questions. For example: What are the requirements and skills that students and professionals in Statistics are requested on employment notices published?
For the first answer, we could see the Word cloud of the complete database in order to discover the main requirements founded.

Fig. 4. WordCloud considering the entire Corpus (Total: 2,809 job postings)

As observed, the most prevalent and relevant terms in the job appear larger. That is, in a high percentage of postings, these words appeared which leads us to believe that one of the first things required of a statistic is the experience ("Experiencia"). We can see other some basic and generic skills, data analysis and information management ("datos", "análisis", "información" y "manejo"). Then, some other words make references to specific skills, such as SPSS or Excel. So, it is necessary to use clustering techniques, since there are several groups of words representing different capabilities related to statistical profiles.

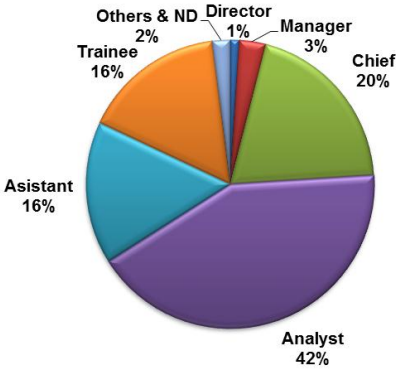


Fig. 5. Distribution of job postings for level (Total: 2,809 job postings)

It's clear that analysts' position dominates, because as we said, the job postings correspond to basic or intermediate positions.

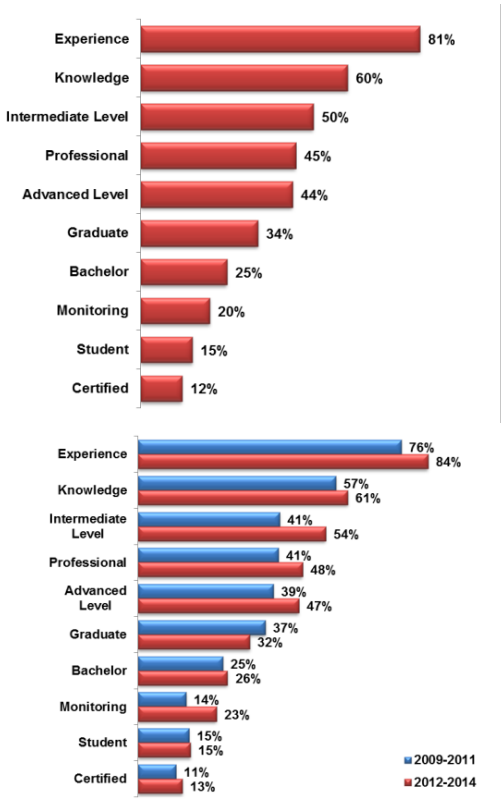


Fig. 6. Job postings distribution by Requirements and period

It is remarkable that 81% of job postings mention the word “Experience” in them. It means that this is one of the most important requirements (along with knowledge or intermediate and advanced levels). Furthermore, they have experienced increasing importance in recent years.

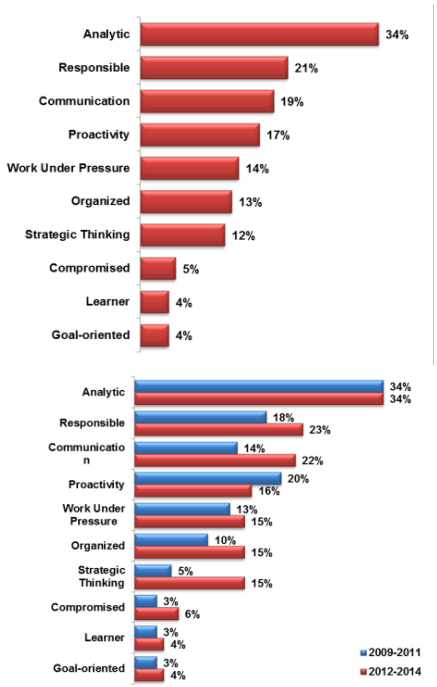
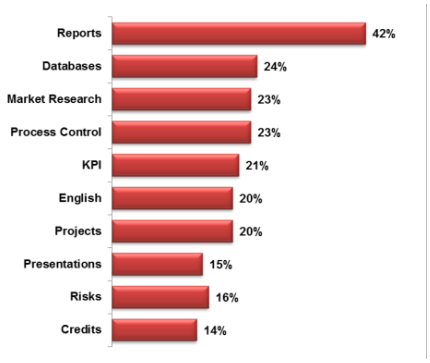


Fig. 7. Job postings distribution by Competencies and period

As for the Competencies, we highlight the character or analytical profile along with other basic skills in business such as responsibility and communication skills. The increase of good communication, responsibility and strategic thinking is valuable. Clearly, the organizations seek Statisticians that are not only good at technical level, but also have the ability to think about the best solution for the organization as a whole.



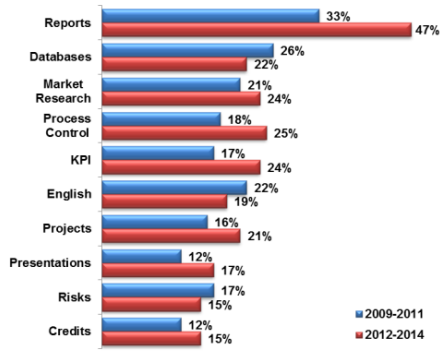


Fig. 8. Job postings distribution by required background and period

About the background required, it weighs heavily reporting tasks or report writing (24%). One in four job postings, contains the term "database" which makes clear that the SQL language has become very important in Lima. Not just someone who can get statistics or models is needed, organizations valued professionals whose can extract themselves from the data sources. Other tasks are in high demand as Process Control or Indicators Development.

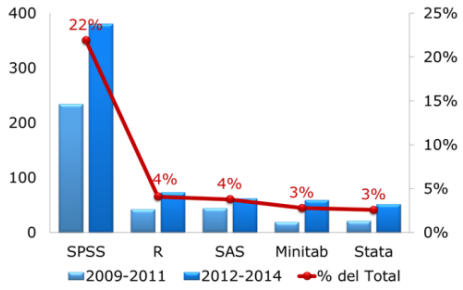


Fig. 9. Most required Statistical Software

The importance of SPSS in the area of Lima is also clear growth in recent years (almost doubling its appearance in the ads). Others such as R or SAS are still not much required; maybe because the cost of acquisition or the time required learning the software (SPSS is easier).

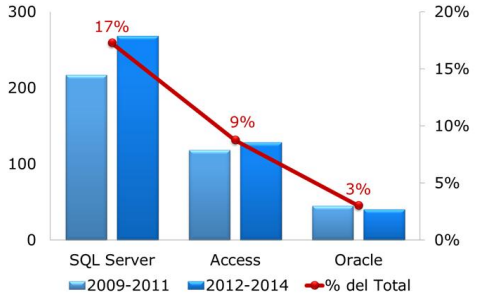


Fig. 10. Most required Database Management Software

Regarding the database software, SQL Server predominates over Access or Oracle.

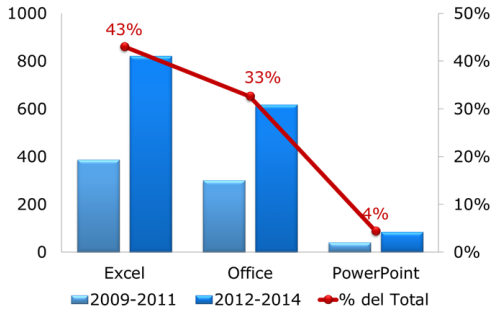


Fig. 11. Most required Office Software

Notice the importance of Excel (appears in four out of ten job postings) and its great increase in recent years.

Finally, it is important to determine the existence of specialization profiles, segments that meet specific characteristics and are different from others. For this, we use SAS Enterprise Miner to compare the results from four, five and nine segments, we decided to choose five segments because it showed better indicators of distance between clusters and better possibilities of interpretation. The distribution of each segment is shown in the next figure:

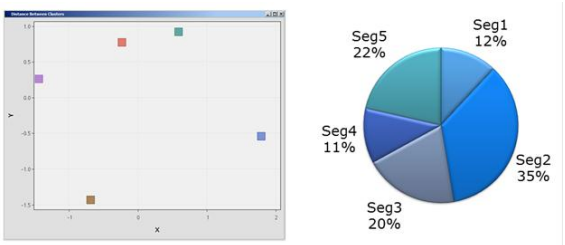


Fig. 12. Term-based Segmentation in SAS Enterprise Miner

After segmenting the messages in these groups, we decided to perform a characterization, that is, find the most common expressions in each cluster, in order to get a better idea of the composition of each segment:

	Seg1 Risk Management	Seg2 Reporting	Seg3 Business Intelligence	Seg4 Trainee	Seg5 Market Research
Knowledge.1	Risk Management	Data Analysis	Databases	Investigate	Investigate
Knowledge. 2	Stocks Management	Comercial	Marketing	Engineering	Techniques Cuantitativ
Profile 1	Analysts	Marketing	Campaign Management	Proactive	Exp. Surveys
Profile 2	Engineers	Data Analysts	Experience in B.I	Graduate	Exp. Marketing
Software 1	SQL	Excel	Excel	SPSS	SPSS
Software 2	SPSS	Office	SQL	Excel	Excel

Fig. 13. Segments Characterization

Through descriptive terms offered by the five clusters finally formed and considering the results of characterization through WordClouds. The following professional profiles were obtained:

Risk managers (Cluster 1): Professionals with experience in portfolio and risk management (both credits and investments), preferably analysts and engineers. They are sued for the financial and banking sector. They were also requested domain mainly SQL and SPSS.

Fig. 14. Word cloud of the Cluster 1

Analysts with reporting tasks (Cluster 2): Analysts with good statistical knowledge required for tasks of reporting and report writing. Mainly related to the areas of marketing and sales. The most required software is the Office suite, more specifically Excel.

Fig. 15. Word cloud of the Cluster 2

Business Intelligence Professionals (Cluster 3): Profiles that manage and analyze databases generally related to marketing and related areas (customers, sales, campaigns). They were also asked experience in campaign management and business intelligence. In software they are required Excel and SQL.

Fig. 16. Word cloud of the Cluster 3

Students or graduates in trainee programs (Cluster 4): Young graduates who are at the end of its cycle of studies (generally engineering) with knowledge of analysis tools and required to be proactive. They are required to dominate Excel and SPSS.

Fig. 17. Word cloud of the Cluster 4

Market researchers (Cluster 5): Professionals in the field of market research (both quantitative and qualitative analysis). They were also required experience in processing and analysis of surveys and marketing knowledge (for research applications). They are required Excel and SPSS too.

Fig. 18. Word cloud of the Cluster 5

These are the profiles we wanted to find, as we have seen, each implies that the professional should have sought some proper statistics to job in question features.

5 Conclusions

According to the results, we can conclude that Statisticians have relative success in Lima. In addition, we have obtained the following conclusions:

1. The main goal (to identify key competencies and requirements) has been successfully achieved. It was possible to detect the main (technical and personal) requirements that often companies require in their job requirements. And due to the temporary separation into two periods, we also found interesting differences about the change in the demand of these requirements.

2. The second one (identification of professional profiles), has also been achieved. We have identified five types of professionals, each group are different from the rest and we have characterized them accurately and in a very clear way.

3. The results obtained in this analysis, may be useful for three agents who are involved in the labor market: companies, potential workers (statisticians) and educational institutions:

- Business: Companies can improve their job postings, making easy the contact with the wanted profiles. In the other hand, they could obtain certain advantages in areas such as employee training, based on the specific profiles founded.

- Statisticians: This analysis would be helpful for them, in order to improve the CV writing, increasing their chances to obtain a good employment opportunity. They can also focus their training in the same direction as do the requirements of companies.

- Education: Universities, training centers and other institutions can adjust their academic offer, in order to meet the needs of the market.

References

[1] AMSTAT (2015). "Statistics is the fastest-growing undergraduate degree". [Consulted: February 3, 2015]. Available in: http://bit.ly/1uvCn4F

[2] Cajachahua, L. (2008). “Estadísticos de Perú”. Blog de empleo y prácticas. [Consulted: February 15, 2015]. Available in: http://bit.ly/1FZVfuV

[3] Cox, A., and Corral, S. (2013). "Evolving Academic Library Specialties". Journal of the American Society for Information Science and Technology. 64 (8): 1526-1542.

[4] Domenica, F., Mastrangelo, M., and Sarlo, S. (2012). "Text Clustering Based on Centrality Measures: An Application on Job Advertisements". [Consulted: June 1, 2015]. Available in: http://bit.ly/1HO6uVv

[5] ElPais.com (2014). "Las carreras con mayor tasa de empleo". [Accessed: October 29, 2014]. Available in: http://bit.ly/1rSot5P

[6] ElPais.com (2015). "¿Cuáles son los estudios con menos paro? ¿Y los que más tienen?" [Consulted: May 7, 2015]. Available in: http://bit.ly/1Jt25K3

[7] Han, J., and Kamber, M. (2001). Data Mining: Concepts and Techniques. Morgan Kaufmann.

[8] IPEA (2014). Radar: Technology, produção and Foreign Trade (2013) 27 Institute of Applied Economic Research. Setoriais Diretoria of Studies and Policies, of Inovação, Regulação and Infrastructure. [Consulted: June 1, 2015]. Available in: http://bit.ly/1SZHL9j

[9] LinkedIn (2014). "The 25 Hottest People Skills That Got Hired in 2014". [Consulted: December 17, 2015]. Available in: http://linkd.in/1x0LQBT

[10] LinkedIn (2014). "Top 10 Job Titles That Did not Exist 5 Years Ago". [Consulted: June 1, 2015]. Available in: http://linkd.in/KtpUbI

[11] Merca20.com (2014). " Infografía: 10 populares empleos que no existían hace 5 años". [Consulted: June 1, 2015]. Available in: http://bit.ly/1abEw6c

[12] Parr Rud, O. (2001). "Data Mining Cookbook". John Wiley & Sons, New York, NY.

[13] RPP.com (2015). "Conoce cuáles serán los empleos más demandados en los próximos 10 años". [Consulted: March 4, 2015]. Available in: http://bit.ly/1EYJH7k

[14] Swan, A., and Brown, S. (2008). "The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment on Current Practices and Future Needs". Report to the JISC.

[15] Thompson, Cheryl A., and Craig Willies. (2015). "Data Workforce Needs: Disambiguation of Roles Using Clustering and Topic Modeling". [Consulted: June 1, 2015]. Available in: http://bit.ly/1QaPDpu

[16] Witten, IH, Frank, E., and Hall, MA (2011). "Data mining: Practical machine learning tools and techniques". San Francisco: Morgan Kaufmann. 3rd edition.

[17] National Institutes for Health (2015). PubMed: US National Library of Medicine. [Consulted: June 1, 2015]. Available in: http://1.usa.gov/1brVEaa