EMPLOYEE EXPERIENCE ANALYSIS IN LABOR SOCIAL MEDIA

APPLYING NLP TECHNIQUES

CASQUERO BOTE, MARIA ISABEL
KSCHOOL
Data Science Master (2020-2021)

1 TECHNICAL ENVIRONMENT AND TOOLS

In this section I will describe the technological environment that I have used to develop this project, including operating system, virtual machines, programming language, libraries, etc.

- Linux: operative system (OS) used under Xubuntu flavor, which is a Linux distribution based on Ubuntu. Shell command-line interpreter has been used to interact with the system through scripting language.

- Oracle VM Virtualbox: is an open source hosted hypervisor developed by Oracle. It allows launching a virtual machine with a different operating system than our computer, which will allow us to say, work in a parallel environment and adapted to the needs of the project, which can even have a greater computing capacity.

- Anaconda: open-source most popular data science platform, because hundreds of the most commonly used open-source data science and machine learning languages or packages are automatically installed and ready to be used when you download this platform. I have used the **conda - -version 4.6.8.**

- Jupyter-lab: web-based interactive notebook interface that offers an interactive and easy to use development environment as alternative to the classic IDE, for notebooks, code in different languages, markdown writing, data, etc. It is developed by Project jupyter and allows you to use different jupyter notebooks for the same project, within the same technical environment.

- Markdown: a lightweight markup language created by John Gruber that tries to achieve maximum readability and ease of publication in both its input and output forms. It has been used to write the readme file and for all the text included on Jupyter notebooks.

- HTML: a code language where web scraping comes in. It has been used in the first phase of this project, where I had to capture all the data that has been subsequently exploited. The scraping is the act of taking data from the HTML code of websites and collecting it into one place for your personal use. The HTML code then gets broken down and separated according to the information that we want and require.

- Python: open-source programming language widely used in data science due to the readability of its code, which can also be used for software development. I have used the **python - -version 3.7.1**

  The *packages or libraries* that have been used in this project are:

  - Web scraping:
    - selenium
    - bs4 (beautifulsoup)
    - requests
  - Data manipulation and analysis:
    - numpy
    - pandas

- Date:
    - datetime
    - time
- Serialization
    - pickle
- EDA Visualization:
    - matplotlib (pyplot, ticker)
    - seaborn
    - wordcloud (WordCloud, STOPWORDS, ImageColorGenerator)
    - pillow (Image) - to generate a wordcloud image
- Natural Language Processing (NLP):
    - Pre-processing
        - Re - regular expressions (a way to search patterns)
        - String
        - NLTK: tokenize, corpus (words, stopwords)
        - Gensim: gensim.utils (simple_preprocess)
    - Advanced pre-processing
        - N-grams
            - Gensim: models (Phrases), models.phrases (Phraser)
        - Stemming & Lemmatization
            - NLTK: tokenize, corpus (wordnet), stem (WordNetLemmatizer, PorterStemmer)
        - Document-Term Matrix
            - Gensim: corpora
    - Sentiment Analysis (NLU)
        - NLTK: sentiment.vader (SentimentIntensityAnalyzer), sentiment.util
    - Topic Modeling (NLU)
        - LDA model training
            - Gensim: models.LdaMulticore
            - pprint - to print keyword in topics
        - LDA metric: coherence score

        - Gensim: models.coherencemodel (CoherenceModel)
    - Parameters optimization: hyperparameters metric
        - Gensim: models.ldamodel.LdaModel
    - LDA visualization
        - pyLDAvis: gensim_models
- Tableau: data analysis and interactive visualization software focused on business intelligence, belonging to an American company with the same name. I have used to make de exploratory or descriptive analysis of some information on this project. It has helped me understand what data I have and answer some of the questions raised in this project, creating interactive dashboards that allow an easy and agile display of information.

2 INTRODUCTION

Historically, in the people area of any organization, decision making has always been based on intuition. By leaders with extensive experience and knowledge in the area, of course, but without data to support their decision.

Advanced analytics applied to people changes this paradigm completely. In other words, it goes **from intuition-based decision making to evidence-based decision making**, that means, based on insights provided by data analysis.

Within this new paradigm, the first problem that we find in the area is having the necessary data. Over time, numerous software packages have emerged which have facilitated the processes of personnel management, payroll, compensation and benefits, controlling, etc (SAP HCM, Oracle People Soft, Sage, Workday, etc). Subsequently, software applications related to the management of recruitment processes have emerged, called ATS (Applicant Tracking Systems), such as Greenhouse, Cornerstone or Bizneo more recently; as well as other more complete software that facilitate the recruitment, onboarding, performance and even analytics processes, and they can be integrated with their own HCRM such as SAP SuccessFactors or Workday itself.

These software packages allow obtaining information in a structured way and even a descriptive analysis derived from it, as I mentioned. The problem comes when we want to go further and solve more complex business questions than the typical KPIs that are traditionally handled in the area (turnover, attrition, rotation, etc). The problem comes when we want to go further and solve more complex business issues than the typical KPIs that are traditionally handled in the area (turnover, desertion, turnover, etc). These problems require unstructured information, and obtaining, cleaning and analyzing it requires more complex data sources, technologies and skills.

We are going to put ourselves in situation, explaining a real and current problem next.

Let's think about how digital transformation is impacting companies. Automation is making thousands of jobs disappear and transforming thousands of others. To deal with these changes, highly qualified personnel with the required skills. On the one hand, they will have to carry out reskilling and upskilling processes for a large part of their workforce and, on the other hand, they will have to go to the market to hire those profiles that have technical skills that they do not have internally or that they have but barely.

Regarding this, the problem arises because what we are experiencing is like a new revolution (technological, in this case). Therefore, it is not an isolated problem that one or several companies have to face, all companies have to face it. In other words, the demand for these highly qualified technical profiles, which are already scarce in the market, is tremendously high. Otherwise, the offer is very small, because this circumstance or scarcity means that all these profiles are currently working and with good working conditions. And this is like the law of supply and demand of all life: as demand rises, supply decreases and prices rise (salary, benefits, etc).

In summary, the consequence of this paradigm is a huge problem for talent departments that are in charge of recruiting and retaining staff. As we advance in the digital transformation, because the need for it is more latent for the companies, it is more difficult to fill a vacancy and retain an employee with these highly qualified profiles. The fight between companies is getting higher, competing with each other for the same profiles.

This forces companies to go further, to deeply analyze the personality, preferences, needs, etc., of these profiles, as well as what the competition offers, to try to improve their recruitment and retention processes. Regarding this, analyzing the experience of employees throughout their life cycle in the company offers us very valuable information.

Certains techniques have traditionally been used to analyze the experience of employees. The techniques most commonly used are the personality tests during recruitment, questionnaires after onboarding, the usual annual speakup surveys launched by companies to their entire workforce, or the usual exit interviews, among others.

All these techniques have a part with closed questions that allow to get structured information automatically, and another part with open questions through which we get 'text' as information. This 'text' is traditionally manually parsed by the appropriate employee within the HR team itself. But let's think that it is a tedious and subjective task, especially if we think in companies with tens or hundreds of thousands of workers. For this reason, it would be very useful to have techniques to overcome this *subjectivity* (analyzing the information based on what the data determines instead of an opinion) and this *tediousness* (automating the information collection and analysis process).

In this project I am going to focus on correcting subjectivity, as a starting point, through the descriptive analysis of the information. Being text, it is necessary to use natural language processing techniques.

Regarding solving the tediousness through automation, this can be the future step for the continuity and improvement of the project.

3 BUSINESS QUESTIONS AND PROJECT GOALS

In relation to carrying out an analysis of the experience of the employees with the final goal of improving our recruitment and retention processes, we are interested not only in knowing their opinions about their experience in our company, but also the experience they have had in other companies of the competition. With the aim of knowing what the competition offers to get ahead of it in the fight for these highly qualified profiles.

Every analytics project starts with business questions to answer. In this case and paradigm, the questions that arise are the following.

Problem 1: I need to know the opinion of employees about different competing companies that 'fight' for IT profiles.

Questions:

- What feeling do they have about their experience in competing companies? is it positive, negative or indifferent?
- What is the sentiment that prevails in the opinions about the different companies (comparing them)?

Problem 2: I need to know what aspects have influence or impact in their experiences, in order to they can determine if these experiences have been good, bad or indifferent.

Questions:

- What factors are most influential in their labor satisfaction or dissatisfaction?
- What topics are derived from their opinions about their experience?
- Which of these topics stand out above the rest because they are more named and therefore have more relevance for the employee?

Problem 3: the interesting thing is to make a comparison between different sectors or groups of companies, to determine which of these are more attractive for the recruitment of this type of profiles.

Questions:

- Are the competitors´ employees more or less satisfied?
- What are the sectors and companies with the best employee experience and satisfaction? Why? What factors influence it?

Regarding to solve these questions, in summary, the main goals in this project will be the following:

- **Descriptive analysis** using NLP techniques, to find out in which sectors and companies a better employee experience is offered and why.
- **Sentiment analysis** (NLU) to score and label the employees sentiments.
- **Topic modeling** (NLU) to discover which are the most relevant topics for employees in relation to their job satisfaction and when determining if their experience has been positive, negative or indifferent.

4 OBTAINING THE RAW DATA

In the introduction, I mentioned the tools traditionally used by human resources teams to analyze the employee experience in the company. Among these tools, the surveys that companies launch to their employees are widely used, especially at three key moments in their life cycle in the company: after their incorporation (at the end of the recruitment and onboarding process), annually throughout their career in the company; and after the end of their employment relationship.

These surveys are usually sent to the employees digitally, by email or through internal platforms. On the one hand, they have a part of closed questions that allow us to obtain structured information and whose quantification is automated by the platform itself, which guarantees agility and objectivity in the data, but less depth in terms of feedback and therefore in terms of its utility. On the other hand, they have another section of open questions that allow us to obtain unstructured information in text format that is not analyzed by the platforms and must be done manually, with the tediousness and subjectivity that it entails and that I commented on in the introduction.

Some examples of surveys from various companies, to understand the structure I have mentioned, are attached in the annex section of this report [12.A].

Due to the impossibility of obtaining these internal tools and its results from my own company for data protection, I considered capturing external data referring to the opinions of employees in their companies. I decided to get the data from Glassdoor as it is the pioneer and most used platform for writing business reviews.

Regarding to the scraping phase, I have obtained global information on opinions about companies based in Spain and, when making the comparison between companies, I will focus on companies that have a high presence in Barcelona since it is the region where there is more competition to recruit and retain these profiles.

On the other hand, when analyzing different competing companies, I will focus on four groups of companies and, within them, in companies that have great strength within the technology market: due to their capabilities and expertise to generate business and to hire a highly qualified workforce. These groups are:

- IT Consulting
- Business consulting
- Own IT product (startups unicorns)
- Final companies from various sectors (banking, telecommunications, pharma, marketplace, tourism, retail, videogames, automotive, etc.)

I will analyze global information of employees who are or have been working within companies belonging to these four groups. And after that, I will focus on the first three groups to make a comparison between companies.

4.1. DATA SOURCING

The data related to the opinions of employees about their experience and job satisfaction in companies across Spain has been obtained from the 'Glassdoor' website as I have said, originally dedicated to reviews of companies at a global level, but which also works as an employment portal.

Let’s have a look to the website: https://www.glassdoor.es/index.htm

4.2. DATA SCRAPING

To get the data from the Glassdoor website, it is necessary to do web scrapping. Regarding this process, I have used the libraries specified below:

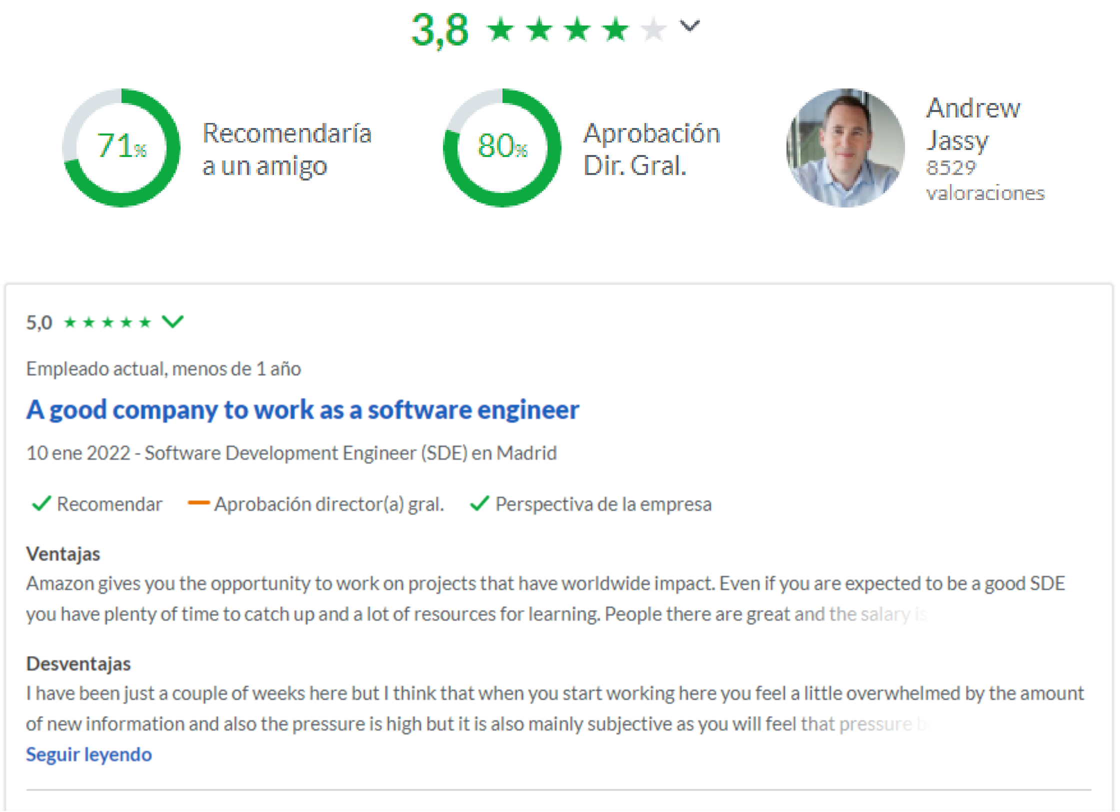
- Selenium: to automate the access into the websites

  To work with selenium, I have carried out the following steps:

  1. Install Selenium library on Jupyter Notebooks
  2. Download webdriver for firefox into the local disk from the following source: https://github.com/mozilla/geckodriver/releases
  3. Import the library

- Request: to get requests to the different websites
- Beautiful Soup: to download the content in html and create an object tree
- Pandas: to create tables with the data obtained
- Numpy: to create dataframes from tables created with numpy

Let's take a look at how reviews are structured on Glassdoor, taking a random Amazon review as an example:



3,8 ★★★★☆

71% Recomendaría a un amigo

80% Aprobación Dir. Gral.

Andrew Jassy
8529 valoraciones

5,0 ★★★★★

Empleado actual, menos de 1 año

A good company to work as a software engineer

10 ene 2022 - Software Development Engineer (SDE) en Madrid

✓ Recomendar — Aprobación director(a) gral. ✓ Perspectiva de la empresa

Ventajas

Amazon gives you the opportunity to work on projects that have worldwide impact. Even if you are expected to be a good SDE you have plenty of time to catch up and a lot of resources for learning. People there are great and the salary is

Desventajas

I have been just a couple of weeks here but I think that when you start working here you feel a little overwhelmed by the amount of new information and also the pressure is high but it is also mainly subjective as you will feel that pressure b

Seguir leyendo

About the review structure that we can observe, in my purpose it would be interesting to capture the following:

- Rating: overall score given to the company
- Title: a summary of the opinion given by the employee
- Date: when the review is sent
- Location: company headquarters where the reviewing employee is located
- Job title: role that the reviewing employee has in the evaluated company
- Advantages: it is assumed that in this section the employee makes positive comments about his satisfaction and experience in the evaluated company. It will be later compiled in the “Pros” feature.
- Disadvantages: it is assumed that in this section the employee makes negative comments about his satisfaction and experience in the evaluated company. It will be later compiled in the “Contras” feature.

In the scraping, the tasks to be done are the following:

- Getting reviews and creating a dataframe per company (all this data has been exported into .csv files and provided on the repository)
- Merging it all into an unique dataframe, to work more easily
- Creating an unified .csv file, which is also provided

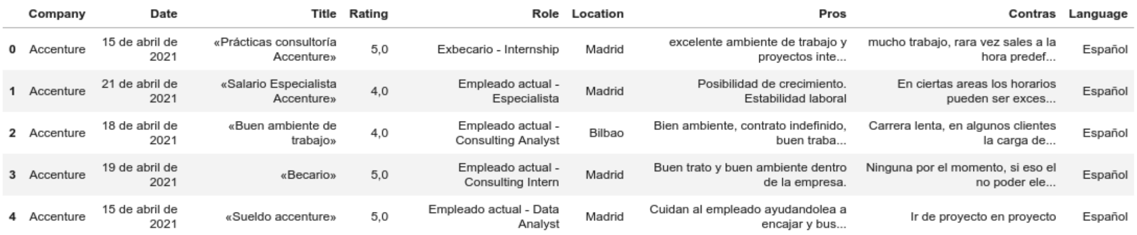
During this process, the issues I have found are specified below:

- Using network headers path to log in into the website
- Checking if all the pages are allowed to scrap or if Glassdoor is blocking any of them
- I had to use Selenium to automate access, click 'cookies' and 'log in' buttons, enter credentials, etc
- The reviews are not shown by location for all companies, which would facilitate the scrapping, but I have had to do the scrapping by company and by language. That is why I had to make a dataframe for each company separately, and the merge all into a single one.

4.3. DESCRIPTION OF INITIAL DATASET

As result of the scraping process, the final .csv file is ***'Companies_reviews.csv'***, which is provided. It is the initial data I have to work on this project.

Let's have a look to this dataset:



	Company	Date	Title	Rating	Role	Location	Pros	Contras	Language
0	Accenture	15 de abril de 2021	«Prácticas consultoría Accenture»	5,0	Exbecario - Internship	Madrid	excelente ambiente de trabajo y proyectos inte...	mucho trabajo, rara vez sales a la hora predef...	Español
1	Accenture	21 de abril de 2021	«Salario Especialista Accenture»	4,0	Empleado actual - Especialista	Madrid	Posibilidad de crecimiento. Estabilidad laboral	En ciertas areas los horarios pueden ser exces...	Español
2	Accenture	18 de abril de 2021	«Buen ambiente de trabajo»	4,0	Empleado actual - Consulting Analyst	Bilbao	Bien ambiente, contrato indefinido, buen traba...	Carrera lenta, en algunos clientes la carga de...	Español
3	Accenture	19 de abril de 2021	«Becario»	5,0	Empleado actual - Consulting Intern	Madrid	Buen trato y buen ambiente dentro de la empresa.	Ninguna por el momento, si eso el no poder ele...	Español
4	Accenture	15 de abril de 2021	«Sueldo accenture»	5,0	Empleado actual - Data Analyst	Madrid	Cuidan al empleado ayudandolea a encajar y bus...	Ir de proyecto en proyecto	Español

- Number of reviews (rows): 10488, in Spanish and English
- Reviews of 36 companies in total, the following (alphabetically):
    - Accenture
    - Adevinta
    - Amadeus
    - Amazon
    - Atos
    - BBVA
    - Caixabank
    - Capgemini
    - Criteo
    - Deloitte
    - DXC Technology

- eDreams
- Everis
- EY
- GFT
- Glovo
- HPE
- HP Inc.
- IBM
- Indra
- King
- KPMG
- Mango
- Minsait
- Nestle
- Novartis
- PwC
- Roche
- Banco Santander
- SEAT
- Socialpoint
- Sopra Steria
- Telefonica
- TravelPerk
- Typeform
- Vistaprint

- Features (columns): 9 in total, the following:

  A. Company: the different companies that have been rated and evaluated

  B. Date: when the review was written, with day, month and year. **Data is displayed from 2008 (when Glassdoor started working) to the first quarter of 2021 included (when I finished the scraping).**

  C. Title: a very little summary of the global review (by the employee).

D. Rating: score (1 to 5 stars) that the employee gives to the company that it has been evaluated according to his level of satisfaction.

E. Role: job title that the employee has or has had in the company

F. Location: office location where the employee works or has worked, across Spain regions

G. Pros: supposedly positive opinions that the employee has about his experience in the company

H. Contras: supposedly negative opinions that the employee has about his experience in the company

I. Language: language in which the review was written

5 DATA CLEANING

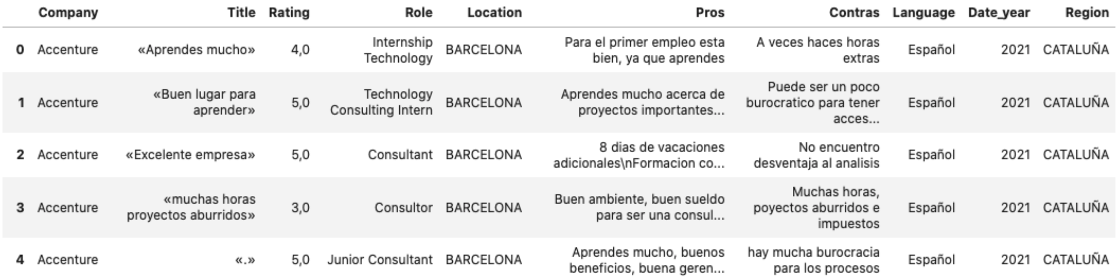
Although during the scraping I have tried to obtain as a result a dataset with the data that I consider useful for my purpose and as clean as possible, it is necessary to proceed with several data cleaning tasks.

These tasks have been the specified below:

- Remove columns that are not necessary for my purpose.
- Remove 'NaN' values.
- Remove words from the 'Date'. Split the date in different columns. Drop day and month. Convert to a datetime object.
- Remove the first part of the text, before the "-", from de original 'Role' column.
- Remove locations outside of Spain (in the scraping, reviews from other countries had sneaked in, which do not interest me for my purpose).
- Join locations by provinces [in 'Location' column].
- Add a new column grouping provinces by regions ['Region' column].
- Create a new .csv file with the dataframe updated.

As result of this initial data cleaning process, the cleaned dataset is collected in the .csv file ***'companies_reviews_v3.csv'***, which is provided.

Let’s have a look to this dataset:



	Company	Title	Rating	Role	Location	Pros	Contras	Language	Date_year	Region
0	Accenture	«Aprendes mucho»	4,0	Internship Technology	BARCELONA	Para el primer empleo esta bien, ya que aprendes	A veces haces horas extras	Español	2021	CATALUÑA
1	Accenture	«Buen lugar para aprender»	5,0	Technology Consulting Intern	BARCELONA	Aprendes mucho acerca de proyectos importantes...	Puede ser un poco burocratico para tener acces...	Español	2021	CATALUÑA
2	Accenture	«Excelente empresa»	5,0	Consultant	BARCELONA	8 dias de vacaciones adicionales\nFormacion co...	No encuentro desventaja al analisis	Español	2021	CATALUÑA
3	Accenture	«muchas horas proyectos aburridos»	3,0	Consultor	BARCELONA	Buen ambiente, buen sueldo para ser una consul...	Muchas horas, poyectos aburridos e impuestos	Español	2021	CATALUÑA
4	Accenture	«.»	5,0	Junior Consultant	BARCELONA	Aprendes mucho, buenos beneficios, buena geren...	hay mucha burocracia para los procesos	Español	2021	CATALUÑA

6 INITIAL EXPLORATORY DATA ANALYSIS (EDA)

After cleaning the main dataset obtained in the scraping process, apart from the python plots that can be seen in the notebook itself, I have used Tableau with the aim of performing the visualization creating an interactive dashboard.

The detailed description of the frontend, the link to the dashboards and how to interact with it, is provided in the section dedicated to this term (8).

As key observations, in summary:

- Of the total reviews, 63% are in Spanish and 37% in English.
- Almost all reviews (all across Spain) are made by employees located in Madrid (52%) and Barcelona (39%).
- The evolution of the total reviews collected is evolving with an upward trend; the growth from 2018 being especially significant.
- The companies with the greatest representativeness of the total reviews of the dataset are those that will be analyzed and compared later in relation to the purpose of this project, belonging to the indicated groups or sectors. Regarding this, Accenture, Deloitte, Everis, Indra, Amazon, etc., are the companies with the greatest representation in the dataset, by that order.

7 NATURAL LANGUAGE PROCESSING (NLP) OR TEXT MINING

Text Mining is the set of techniques and technologies used to explore large amounts of text, automatically or semi-automatically, with the aim of discovering repetitive patterns, trends or rules that explain the behavior of the text. In summary, it is a technology whose aim is the search for knowledge in huge amounts of documents (unstructured information), finding relationships or patterns in the content of these documents.

The text mining process can be divided into three phases:

1. Pre-processing: the texts are transformed into some type of structured representation that facilitates their subsequent analysis. That is, the first step in text mining would be to define and clean the set of documents and avoid their duplication.

From the selected and structured set of documents, we must recognize the ***tokens*** (essential grammatical units), which implies representing the text as a list of words using a vector representation (the most common representation of a token is a word). This process is known as ***tokenization***.

After the tokenization process (it provides us a list of tokens per document or review in our case), which is the minimum necessary to proceed with an analysis that provides an initial MVP, more advanced pre-processing tasks can be carried out, such as create ***n-grams*** and, after that, the ***stemming*** and ***lemmatization*** processes. Both processes will be explained later.

As a result of this pre-processing phase, we will have a collection of documents tokenized and compiled into what is known as a ***corpus*** or ***Document-Term Matrix.***

2. Discovery: step of analysis to discover interesting knowledge or patterns.

3. Visualization: users can observe and explore the results in a simple and agile way.

About its application, text mining can be applied in different areas, such as:

- Information extraction
- Sentiment analysis or opinion mining
- Document classification
- Preparation of summaries
- Text generation

In summary, it is very useful for all companies, administrations or organizations in general, which, due to the characteristics of their operation, composition and activities, generate a large number of documents and are interested in obtaining information from all this high volume of data. It can help them to get to know their clients, users or employees better (as in the case of this project) as well as to know their habits, opinions or preferences; in order to find behavior patterns that help in decision making and in the application of strategies.

7.1 DATA PREPROCESSING

When we deal with numerical data, the cleaning process involves removing null values, duplicate data, dealing with outliers, etc, depending on the initial data we have. In other words, as far as this project is concerned, this has been reflected in the cleaning tasks performed on the raw data obtained from the initial scraping of the Glassdoor information to create the initial dataset.

On the other hand, when we deal with text data, we must additionally perform other cleaning tasks known as "text pre-processing techniques”.

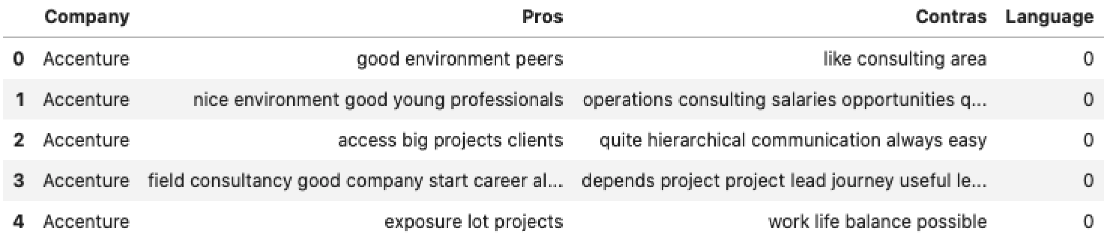
Within this pre-processing step, there are common steps to always take to start from a minimum viable product (MVP) and, from there, there are other techniques to apply to iterate and improve the results which I will apply later.

In this phase, I will apply the common steps for pre-processing **before tokenization**:

- Convert to lowercase
- Remove punctuation
- Remove 'stop words'
- Remove numerical data
- Remove spaces
- Tokenize text

The libraries used in this pre-processing phase are indicated in the first section on "technical environment and tools" (1).

As a result, we get a type of data format known as "**corpus**", a collection of texts collected in a dataframe. Example of this corpus (only with English reviews) would be:



	Company	Pros	Contras	Language
0	Accenture	good environment peers	like consulting area	0
1	Accenture	nice environment good young professionals	operations consulting salaries opportunities q...	0
2	Accenture	access big projects clients	quite hierarchical communication always easy	0
3	Accenture	field consultancy good company start career al...	depends project project lead journey useful le...	0
4	Accenture	exposure lot projects	work life balance possible	0

This corpus resulting from the initial preprocessing is compiled into two datasets, which are provided:

- ***'Reviews_preprocessing.csv'***: with all the reviews, in Spanish and English, that will be used to provide the minimum viable product (MVP).
- ***'Reviews_preprocessing_eng.csv'***: with only the English reviews, that will be used to improve the 'product' delivered. That means, for the advanced preprocessing, to improve the EDA analysis, to perform the LDA model and to make the comparison between companies.

7.2 MINIMUM VALUE PRODUCT (MVP): Delivering an initial product or value

7.2.1 EXPLORATORY DATA ANALYSIS (EDA): WORD COUNTS OR TOP WORDS

After the cleaning or text preprocessing phase, in which we have put the data into a standard format, the next step will be to take a look at the text data to determine if it makes sense, before proceeding to use NLU techniques or to train a model.

The idea is the same as when working with numerical data, exploring the data to find obvious patterns before identifying hidden patterns with machine learning techniques.

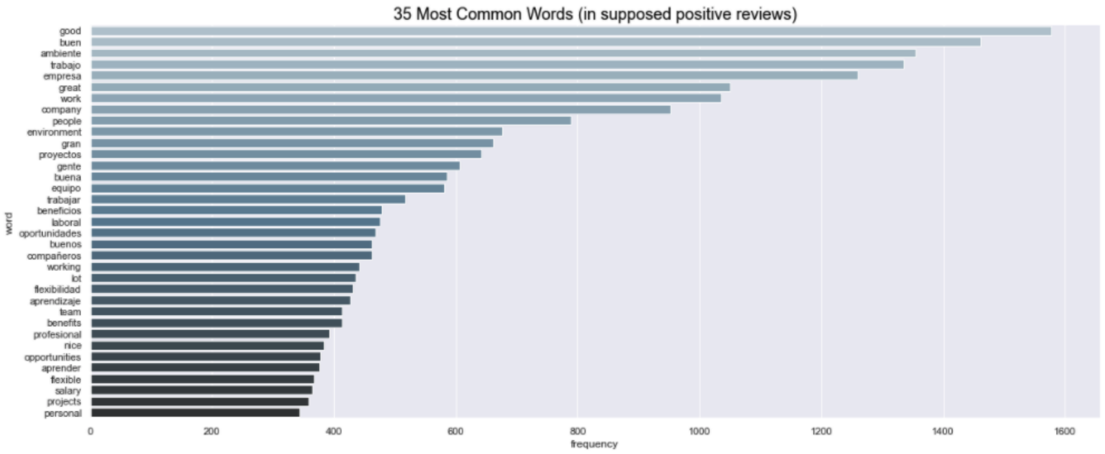
In this phase, in order to have an initial exploration of the words in the data set, I have proceeded with the ***'word counts'***, that is, I have extracted the most common words or ***'top words'*** from the entire corpus (considering the reviews in Spanish and English), but I have extracted the 'top words' for each feature separately ('Pros' and 'Contras'), to determine which words are most common in the supposed positive and negative comments.

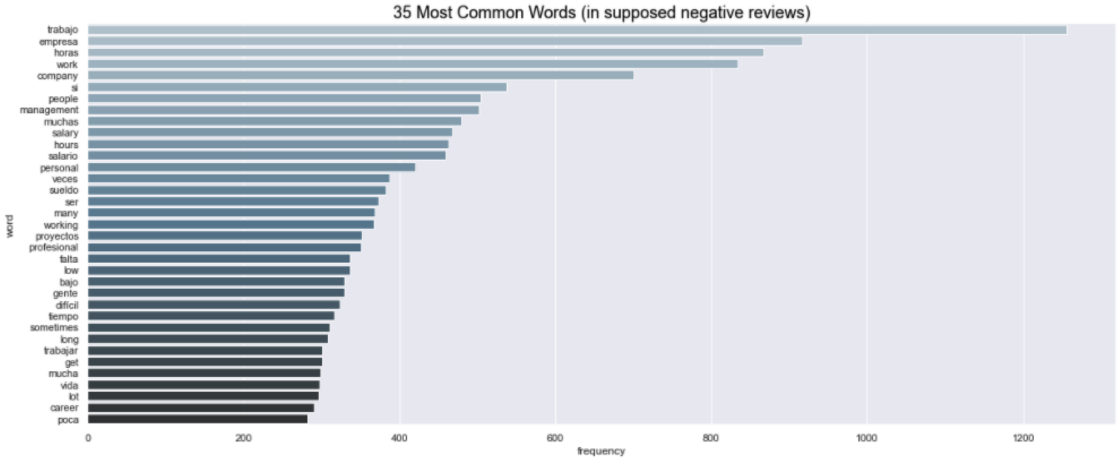
Tasks that I have done:

1. Create a variable with a list of all the words mentioned in the text, one for each feature ('positive_word_list' and ‘negative_word_list').

2. Word counts: count the 100 most repeated or common words of each list, and collect them in two variables ('positive_word_counts' and ‘negative_word_counts').

3. Visualization in 'Bar Charts' of the 35 most repeated words for each feature, using the matplotlib and seaborn libraries.

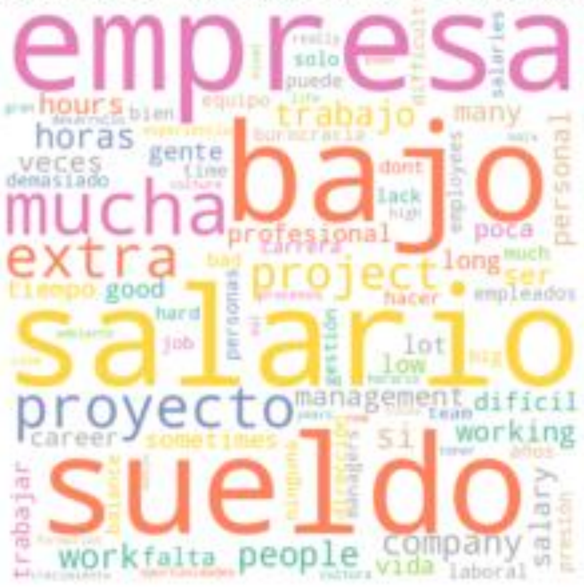




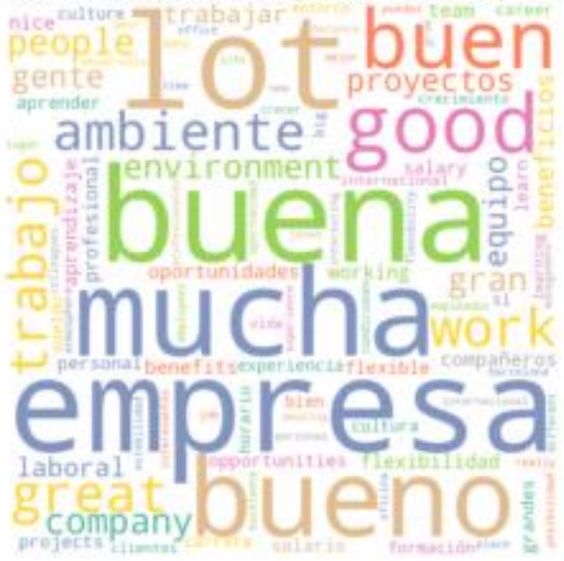
4. Visualization in 'Word Clouds' of the 100 most common and already countable words, for each feature, using the WordCloud (to generate the object) and pillow (to generate the image) libraries, and matplotlib for the display. Steps to be done:

- Extract the review (text document)
- Create and generate a wordcloud object
- Display the generated image

100 Most Common Words (in supposed negative reviews)



100 Most Common Words (in supposed positive reviews)



Later I will get the 'top words' by company to make a comparison, but I will do it in another phase of the project, only considering the reviews in English and gathering both features.

7.2.2 OBSERVATIONS: DOES THE DATA MAKE SENSE?

After this initial exploratory analysis of the text where I have been able to visualize the most common words, it is necessary to ask us:

A) Do the most common words ('top words') for each feature make sense?

It can be seen that the most common words, both in the supposed positive and negative comments ('Pros' and 'Contras' features), make sense.

Key observations:

- Words are shared that are highly representative in both and that are directly related when giving any type of feedback related to satisfaction or work experience, whether it is good, bad or insignificant, especially as regards nouns: job, company, people, projects, benefits, colleagues, learning, benefits, time, office, etc.

- But it is true that with respect to certain nouns there are differences: while "environment", "flexibility", "career", "team", and "opportunities" are more represented in the positive comments (we understand that as positive aspects that create a positive employee experience); others such as "hours", "project" or especially "salary" are more representative of negative comments (we understand that as negative aspects that cause job dissatisfaction and possibly dismissal from the company).

- There is a difference in the representativeness of adverbs and adjectives. For example: good, great, nice, flexible, etc., are more representative of positive comments; while bad, low, many, lacking, difficult, extra, etc., are more representative of negative comments. It makes sense that these adverbs and adjectives refer to important aspects such as the environment, working hours, opportunities for growth, salary, work-life balance, etc.

- Regarding the verbs, we also find differences. While “work” is obviously mentioned in both features, “management” is more representative in negative comments (certainly referring to bosses, managers or the executive team as a reason for discontent), and “learn” is more representative in positive comments (because learning and training is one of the aspects most valued by employees in a positive experience)

B) Should I continue to clean the data through additional preprocessing tasks? maybe removing extra ‘stopwords’, creating n-grams or finding the lemma of the words depending on its meaning and context?

Although the 'top words' make sense, I consider necessary to improve the preprocessing following with more advanced techniques.

Key observations:

- Very represented words are observed that do not contribute anything and that can be added as additional 'stop words' (yes, get, new, better, times, be, do, dont, etc.).

- Regarding adjectives and adverbs, it would be interesting to know what nouns or verbs they are associated with. For example, I can think that "long" refers to "hours", that "flexible" refers to "timetable" or "schedule", or that "social" refers to "benefits". To know this relationship, we can proceed with the 'n-grams'.

- It is observed, especially in relation to adverbs and adjectives, that many of them refer to the same meaning ("much", "lot", "many"; or "low" and "less"). The same issue happens with what the same words provide in plural or singular, or with the same verbs but in different conjugations. To solve this issue, we can carry out the 'lemmatization' process.

In conclusion, the product can be improved or the value added in this project can be increased by improving the preprocessing.

7.2.3 SENTIMENT ANALYSIS (NLU)

Sentiment analysis is the process of determining computationally whether a piece of writing (in this case, a review) is positive, negative, or neutral. It is also known as opinion mining. It allows us to know how people (customers, consumers, employees, etc.) respond to a product, service, experience, brand, etc. In this case, it has allowed me to know what the employees think about their experience in their respective companies, if this experience has been positive, negative or neutral for them.

Therefore, what I have done, have been to label each review as positive, negative or neutral, and by company; to later count the total number of labels for each feeling and company. In this way, I have been able to visualize and determine which companies generate a more positive feeling and which, on the contrary, generate more negative experiences or feelings.

On the other hand, it has also allowed me to determine if the text related to the supposed positive comments in the review structure of Glassdoor ('Pros' feature) actually contains only positive comments; as well as to determine if the text with the supposed negative comments ('Contras' feature) exclusively contain said negative comments.

Regarding to the code, there are many packages available in python which use different methods to do sentiment analysis. Some of the most popular methods and packages are ‘Textblob’ and ‘Vader’ (which I have used it), for example.

'Valence Aware Dictionary and sEntiment Reasoner' is a rule/lexicon-based, open-source sentiment analyzer pre-built library, protected under the MIT license.

I have used NLTK and VADER package for the sentiment analysis of the reviews for the following reasons:

A. Vader performs exceptionally well in the domain of social networks and generalize favorably.

B. When Vader is compared to sophisticated machine learning techniques, its simplicity carries several advantages. First, it is fast and computationally cheap without sacrificing accuracy.

C. The lexical dictionary and rules that Vader uses are accessible as they are not hidden. Therefore, Vader can be easily inspected, understood, extended or modified.

Furthermore, as Vader brings to light his model of lexical dictionary and rules, he makes work in Sentiment Analysis more accessible.

D. Using a lexical dictionary of general sentiments and rules related to grammar and syntax (validated by humans), Vader does not it requires an extensive training set and performs well in a variety of domains.

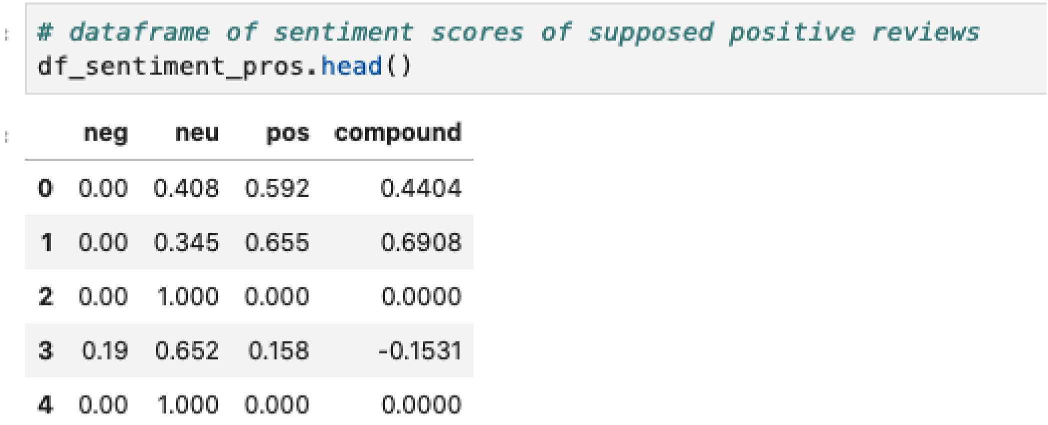
The library is imported as follow:

1. import nltk
2. nltk.download('vader_lexicon')
3. from nltk.sentiment.vader import SentimentIntensityAnalyzer

Vader uses a list of lexical features which are labeled as positive or negative according to their semantic orientation to calculate the text sentiment. Vader sentiment returns the polarity of the word and the probability of a given input sentence to be positive, negative and neutral.

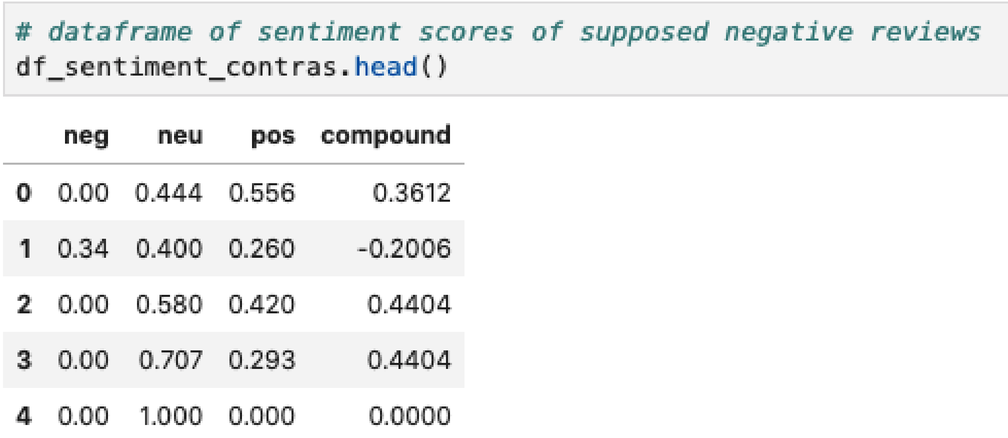
Regarding this and, in summary, the process has been as follows:

1. I have started from the preprocessed dataset with the reviews in English): ‘Reviews_preprocessing_eng.csv’, with a total of 3.691 reviews, each of which has supposed positive comments collected in the 'Pros' feature; and alleged negative comments collected in the 'Contras' feature. Since it is more effective analyzing in English.
2. I have obtained polarity scores for each review and for each feature separately (supposed positive comments vs. supposed negative comments). It gets back a dictionary of different scores. The negative, neutral and positive scores are related, because they all add up to 1 and can’t be negative. But also, it uses the ***'Compound Polarity'*** score metric, that is calculated differently. It’s not just an average, and it can range from -1 to 1. It's a metric that calculates the sum of all the lexicon ratings which have been normalized between -1 (most extreme negative) and +1 (most extreme positive). The results below, in dataframes for both features:



```
# dataframe of sentiment scores of supposed positive reviews
df_sentiment_pros.head()
```

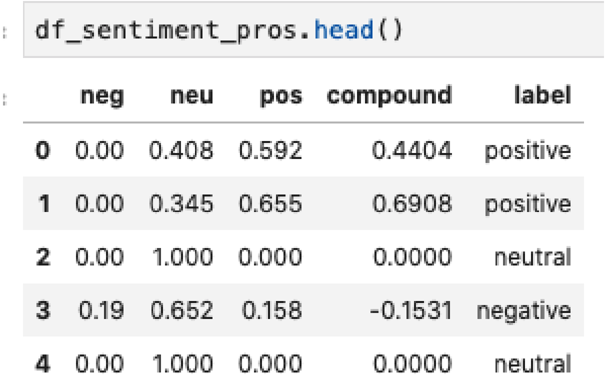
	neg	neu	pos	compound
0	0.00	0.408	0.592	0.4404
1	0.00	0.345	0.655	0.6908
2	0.00	1.000	0.000	0.0000
3	0.19	0.652	0.158	-0.1531
4	0.00	1.000	0.000	0.0000



```
# dataframe of sentiment scores of supposed negative reviews
df_sentiment_contras.head()
```

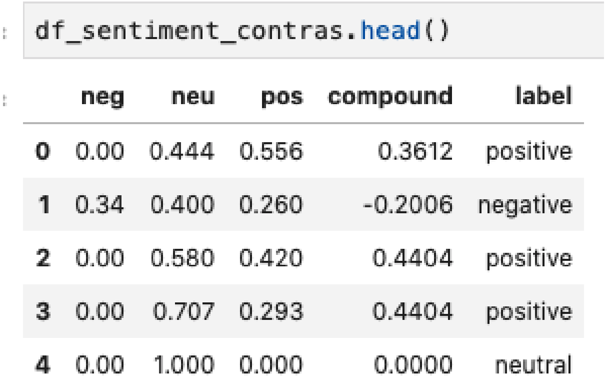
	neg	neu	pos	compound
0	0.00	0.444	0.556	0.3612
1	0.34	0.400	0.260	-0.2006
2	0.00	0.580	0.420	0.4404
3	0.00	0.707	0.293	0.4404
4	0.00	1.000	0.000	0.0000

3. Labeling of reviews based on the value of the compound polarity metric. To label a sentiment as positive, neutral or negative, it has been considered the following:
    - positive sentiment : (compound score > 0)
    - neutral sentiment : (compound score = 0)
    - negative sentiment : (compound score < 0)



```
df_sentiment_pros.head()
```

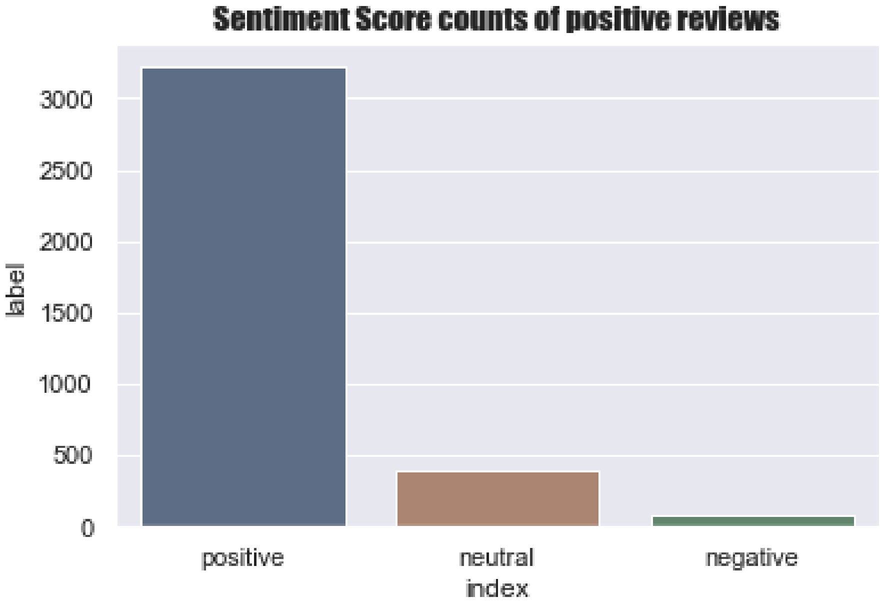
	neg	neu	pos	compound	label
0	0.00	0.408	0.592	0.4404	positive
1	0.00	0.345	0.655	0.6908	positive
2	0.00	1.000	0.000	0.0000	neutral
3	0.19	0.652	0.158	-0.1531	negative
4	0.00	1.000	0.000	0.0000	neutral

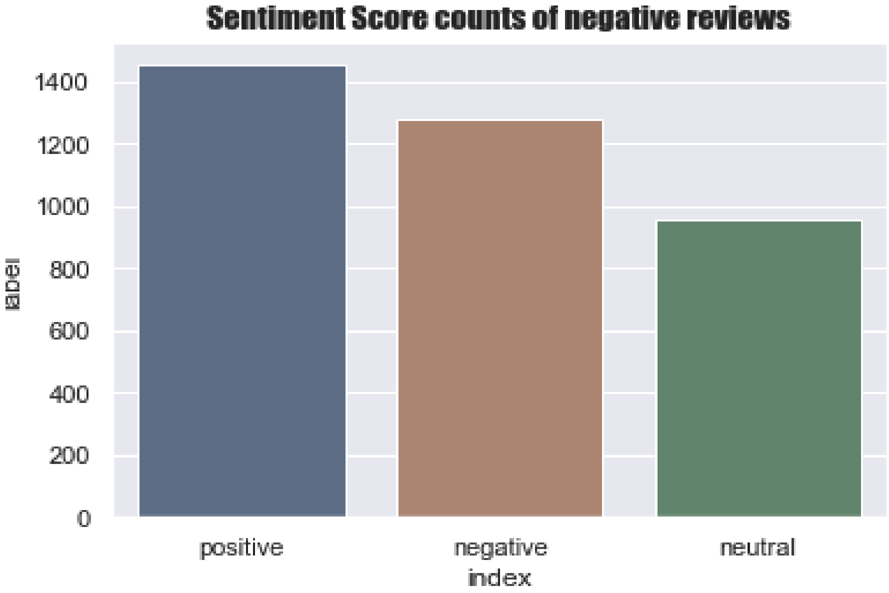


```
df_sentiment_contras.head()
```

	neg	neu	pos	compound	label
0	0.00	0.444	0.556	0.3612	positive
1	0.34	0.400	0.260	-0.2006	negative
2	0.00	0.580	0.420	0.4404	positive
3	0.00	0.707	0.293	0.4404	positive
4	0.00	1.000	0.000	0.0000	neutral

4. Sentiment score counts for both features and plotting:

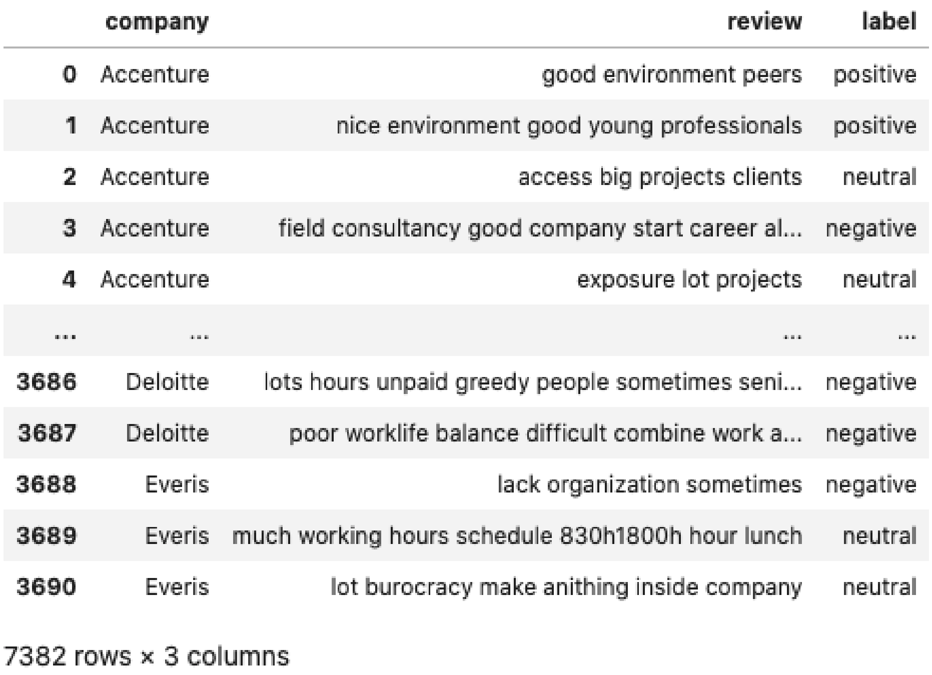




From this distinction, the conclusion is that, while the supposedly positive reviews have certainly been labeled positive for the most part (3,221 positives of a total of 3,691 reviews); the supposedly negative reviews have not. I mean, I observe that regarding the latter, the reality is that more reviews have been labeled with positive sentiment (1,454) than negative (1,277), although with little difference, and there is also a high number of reviews with neutral sentiment (960).

5. After observing what was indicated in the previous point, I have realized that this distinction between features does not contribute much to me, so I have proceeded to unify the features into one, treating them as different reviews, to proceed to unify the

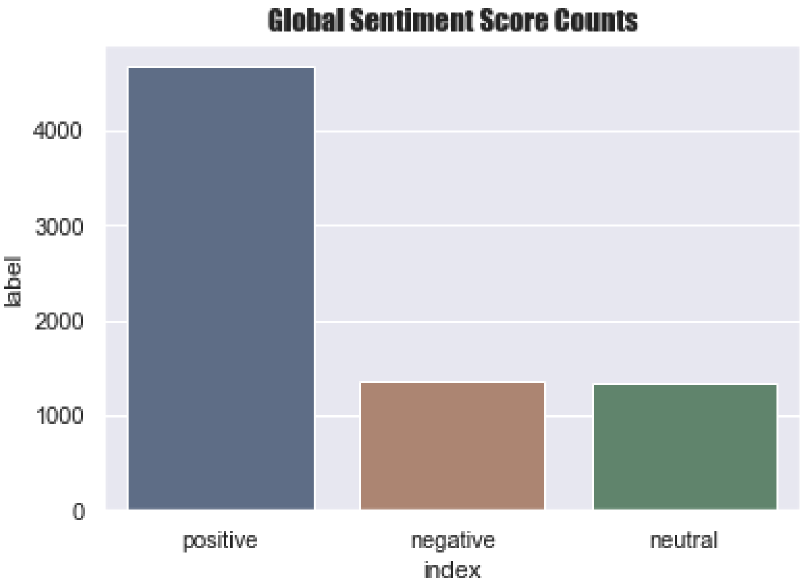
analysis and the results. As result, the new feature is called 'review' and the unified dataframe has a total of 7.382 reviews, all labeled with the type of sentiment.



	company	review	label
0	Accenture	good environment peers	positive
1	Accenture	nice environment good young professionals	positive
2	Accenture	access big projects clients	neutral
3	Accenture	field consultancy good company start career al...	negative
4	Accenture	exposure lot projects	neutral
...	...	...	...
3686	Deloitte	lots hours unpaid greedy people sometimes seni...	negative
3687	Deloitte	poor worklife balance difficult combine work a...	negative
3688	Everis	lack organization sometimes	negative
3689	Everis	much working hours schedule 830h1800h hour lunch	neutral
3690	Everis	lot burocracy make anithing inside company	neutral

7382 rows × 3 columns

6. Sentiment score counts in global.



After counting the sentiment scores for each sentiment per review of the new unified dataframe, a majority of positive reviews can be observed (4,675): 63% of the total of 7,382 reviews as mentioned above. It is also noted that the number of negative and neutral reviews is very similar (1,361 negative and 1,346 neutral).

7. Sentiment score counts per company, collected in a .csv file, and display.

The .csv file is: ***‘sentimentcounts_percompany.csv’***

To visualize this data, on the one hand I have made the plots in python, which can be seen in the delivered notebook; and on the other hand also with Tableau to make the observations in a more agile and interactive way.

The link and way to interact with the frontend in Tableau is explained in the section dedicated to this purpose (8).

As key observations, it is observed that the best positively valued companies have been Typeform, Socialpoint, TravelPerk, Criteo, Vistaprint, King, Amadeus, Adevinta, HP Inc, HPE, in that approximate order. Since they are the ones with the highest proportion of positive comments over the total number of comments collected on each of them in the dataset. With this, we can conclude that startups are the best valued, followed by final companies. While those companies linked to technology or business consulting are the worst rated.

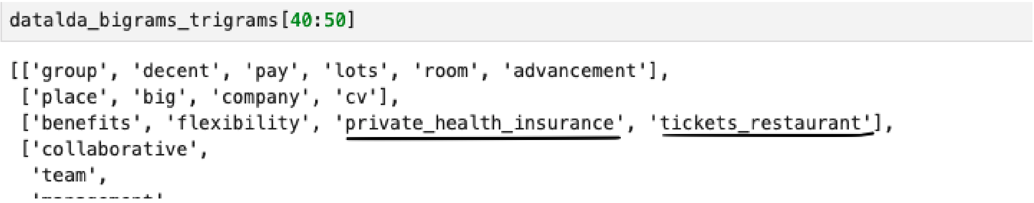
7.3 ADVANCED DATA PREPROCESSING

As I have explained, after the tokenization process (it provides us with a list of tokens per document or review in our case), more advanced pre-processing tasks can be carried out:

- Create *’n-grams’*:

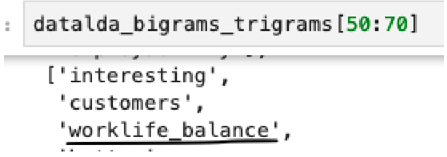
The first thing was to remove additional ‘stop words’ and to create ***n-grams*** (bi-grams, tri-grams...), which consists of selecting 2, 3....n words that usually appear together as the only token.

For example, in this project, I have seen bi-grams such as “tickets restaurant”, “social benefits”, “learning curve”, “career progression”, “flexible schedule”, “worklife balance”, “free coffee”, “long term”, “long hours” or “continuous improvement”; and tri-grams such as “private_health_insurance”.



```
datalda_bigrams_trigrams[40:50]
```

```
[['group', 'decent', 'pay', 'lots', 'room', 'advancement'],
 ['place', 'big', 'company', 'cv'],
 ['benefits', 'flexibility', 'private_health_insurance', 'tickets_restaurant'],
 ['collaborative',
  'team',
```



```
datalda_bigrams_trigrams[50:70]
```

```
 ['interesting',
  'customers',
  'worklife_balance',
```

These *’n-grams’* make sense in texts that are job satisfaction and employee experience reviews, right?

In this process, to create a bigram and trigram models, I have used the *Gensim* library (read section 1).

- *‘Stemming’ & ‘lemmatization’:*

After that, it was what is known as ***stemming and lemmatization*** processes. Thanks to the lemmatization we can relate the inflected or derived words with their canonical form or lemma, just as we would find them in the dictionary.

Both are text normalization techniques.

a. Stemming: words are reduced to their root form, but there are many variations of the same words. It is a technique in which a set of words in a sentence are converted into a sequence to shorten its lookup. In this method, the words that have the same meaning but with some variations according to the context or sentence, are normalized. This process removes redundancy in the data and variations in the same word. As a result, data is filtered which will help in better machine training.

b. Lemmatization*:* words in third person are changed to first person, and verbs in past and future tenses are changed into present. It is the process of finding the lemma of a word depending on its meaning and context. Lemmatization usually refers to the morphological analysis of words, which aims to remove inflectional endings. It helps in returning the base or dictionary form of a word known as the *lemma*.

Text preprocessing includes both, stemming as well as lemmatization. Many people find the two terms confusing. Some treat these as the same, but there is a difference between stemming vs lemmatization. Lemmatization is preferred over the former because:

I. Stemming algorithm works by cutting the suffix from the word. In a broader sense cuts either the beginning or end of the word.

II. Lemmatization is a more powerful operation, and it takes into consideration morphological analysis of the words. It returns the lemma which is the base form of all its inflectional forms. In-depth linguistic knowledge is required to create dictionaries and look for the proper form of the word.

I decided, for this reason, to proceed with the lemmatization process. I have tried to use the Gensim library but I have had issues to import ‘gensim.utils’, so I have used the NLTK library finally, importing ‘WordNetLemmatizer´. The NLTK Lemmatization method is based on WorldNet’s built-in morph function.

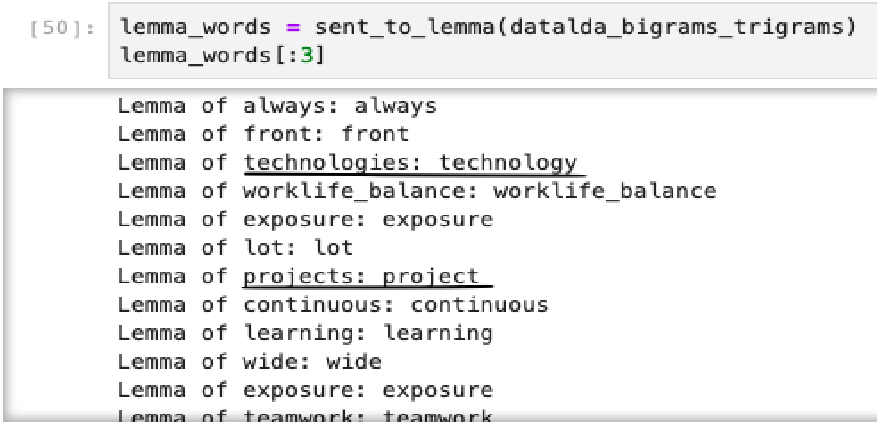
Wordnet Lemmatizer is a publicly available lexical database of over 200 languages that provides semantic relationships between its words. It is one of the earliest and most commonly used lemmatizer technique. Wordnet links words into semantic relations (example: synonyms), and it groups synonyms in the form of synsets (a group of data elements that are semantically equivalent).

To use the Wordnet Lemmatizer, I had to download it from nltk library:

I. import nltk

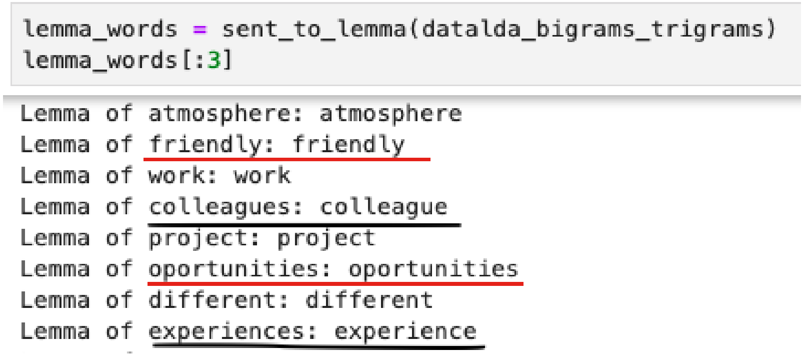
II. nltk.download(‘wordnet’)

III. nltk.download(‘averaged_perceptron_tagger’)



```
[50]: lemma_words = sent_to_lemma(datalda_bigrams_trigrams)
      lemma_words[:3]
```

```
Lemma of always: always
Lemma of front: front
Lemma of technologies: technology
Lemma of worklife_balance: worklife_balance
Lemma of exposure: exposure
Lemma of lot: lot
Lemma of projects: project
Lemma of continuous: continuous
Lemma of learning: learning
Lemma of wide: wide
Lemma of exposure: exposure
Lemma of teamwork: teamwork
```



```
lemma_words = sent_to_lemma(datalda_bigrams_trigrams)
lemma_words[:3]
```

```
Lemma of atmosphere: atmosphere
Lemma of friendly: friendly
Lemma of work: work
Lemma of colleagues: colleague
Lemma of project: project
Lemma of oportunities: oportunities
Lemma of different: different
Lemma of experiences: experience
```

I have found that the result after using Wordnet lemmatizer has not been what I expected. Mainly, the plurals have been transformed into singulars (although it has failed with some tokens), but I also expected to receive the base form of the word as a lemma. In the example below, I expected to receive as return the lemma "friend" (noun) for the word "friendly" (adverb), for example.

This will be reflected in the word count and in the topic modeling. I consider that to improve the project, one of the steps to follow would be to try to improve the preprocessing using the "pos tagging" function from NLTK: *'nltk.pos_tag()'*

Part-of-speech (POS) tagging is a process which refers to categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context, to extract its appropriate lemma.

7.4 TOPIC MODELING (NLU)

Topic modeling has been shown to be useful for automatic topic discovery from a large volume of texts. It allows the organization of information, understanding and extraction of content.

Most prevailing theme models use probabilistic approaches and consider frequency and concurrency to discover the themes of collections of documents. These consider the texts as a mixture of probabilistic topics, where the atopic topic is represented by a probability distribution over the words.

Speaking from a more technical point of view, topic modeling is a type of statistical model to discover the abstract "themes" that occur in a collection of documents. It scans a set of documents (known as a corpus), examines how words and phrases are related, and automatically selects groups of words that best characterize those documents. These sets of words represent a theme (topic).

With the introduction of text mining, research has been conducted to analyze important issues and trends in document collection. Latent Dirichlet Allocation (LDA) for trend analysis in text mining is one of the most accurate trend analysis methods for large texts, but less useful for short texts (such as a “tweet” or phone message, for example).

7.4.1 LATENT DIRICHLET ALLOCATION (LDA)

Latent Dirichlet Allocation (LDA) is a generative, non-parameterized, unsupervised machine learning method introduced as a graphical model by M. Blei, Y. Ng, and I. Jordan in 2003, but discovered by JK Pritchard, M. Stephens and P. Donnelly in the year 2000. Due to its high modularity, it can be easily expanded, giving a lot of interest to your study. It is based on a straightforward probabilistic mathematical concept of Bayesian inference but, despite its strong theory, in the end, it is quite simple to use. Bayesian inference is a way to get more accurate predictions from your data. It's particularly useful when we don't have as much data as we'd like and want to take advantage of every last predictive bit.

An example of using LDA in engineering is automatically classifying documents and estimating their relevance to various topics.

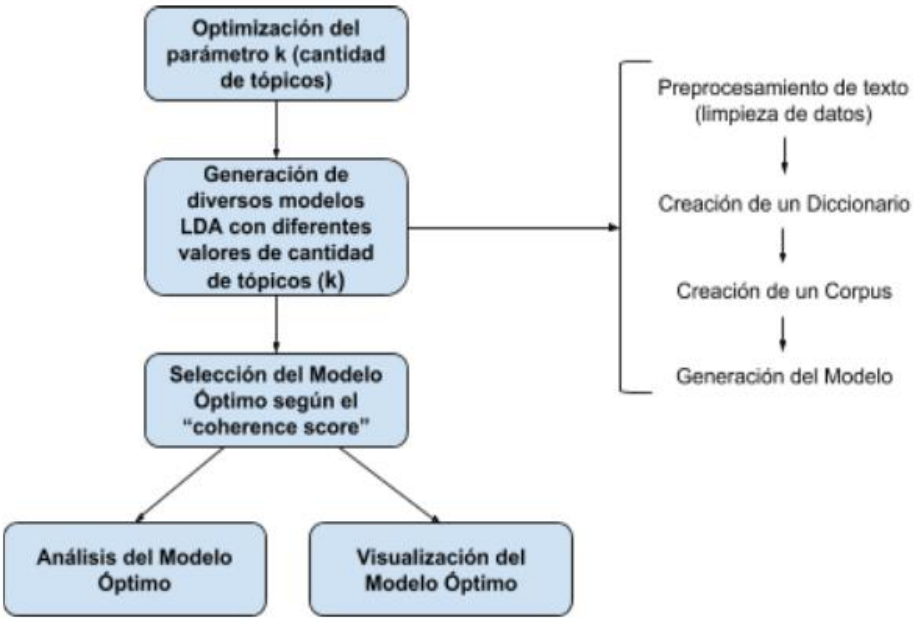
In LDA, each document can be viewed as a mix of various topics where each document is considered to have a set of topics that are assigned to it through LDA. This is identical to probabilistic latent semantic analysis (pLSA), except that in LDA the distribution of topics is assumed to have a *Dirichlet* prior distribution.

Dirichlet's scant background suggests that the documents cover only a small set of topics and that the topics use only a small set of words frequently. In practice, this results in better disambiguation of words and more precise mapping of documents to topics. LDA is a generalization of the *pLSA model*, which is equivalent to LDA under a uniform prior *Dirichlet* distribution.

For example, an LDA model may have themes that can be classified as "cat" and "dog". The former is likely to generate multiple words, such as "milk", "meow", and "kitty", which the viewer can classify and interpret as "cat". Naturally, the word "cat" itself will have a high probability given this topic. The topic "dog", in turn, is likely to generate words like "puppy", "bark" and "bone" that can have a high probability.

Words with no special relevance, such as "the", it will have more or less uniform probability across classes (or can be placed in a separate category). A theme is neither semantically nor epistemologically defined. It is identified on the automatic detection of the probability of simultaneous occurrence of terms. However, a lexical word may appear in various topics with a different probability, with a different typical set of neighbor words for each topic.

The structure or phases of a LDA model, and which I have followed, is reflected on the graph below.


Optimización del parámetro k (cantidad de tópicos)
Generación de diversos modelos LDA con diferentes valores de cantidad de tópicos (k)
Selección del Modelo Óptimo según el "coherence score"
Análisis del Modelo Óptimo
Visualización del Modelo Óptimo
Preprocesamiento de texto (limpieza de datos)
Creación de un Diccionario
Creación de un Corpus
Generación del Modelo

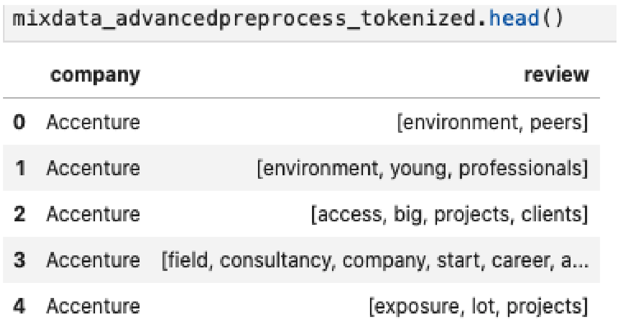
Starting from the data collected in the .csv file *'mixdata_advancedpreprocess_tokenized.csv'*, I have used the Python **Gensim** library to apply this model (check section 1).

This dataset collects only the English reviews, with the features separated and joined, as result of advanced preprocessing and already tokenized. There are a total of 7.382 reviews.

7.4.2 GLOBAL ANALYSIS

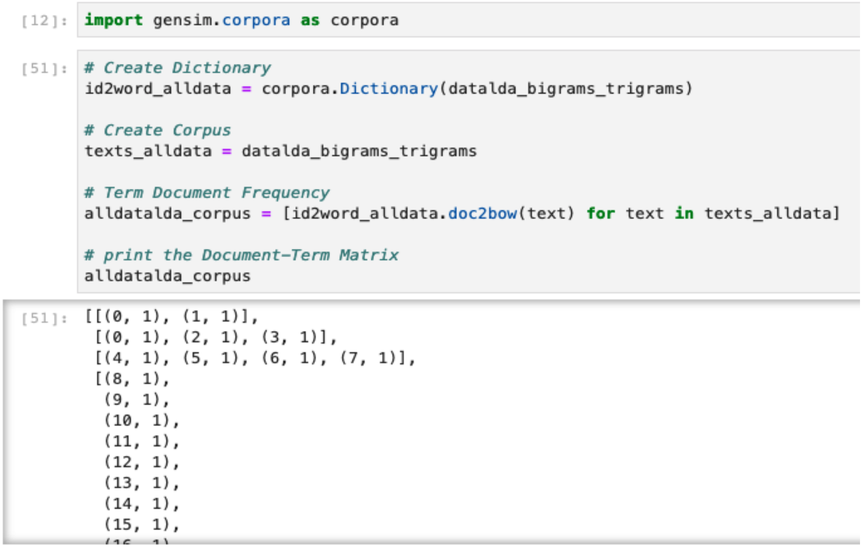
7.4.2.1 Creation of a Document-Term Matrix (DTM) as a corpus

I had to convert the tokenized object already created in a vocabulary dictionary and into a Document-term matrix as a corpus. It represents documents vectors in matrix form in which the rows correspond to the terms in the document, columns correspond to the documents in the corpus and cells correspond to the weights or frequency of the terms in a document.



```
mixdata_advancedpreprocess_tokenized.head()
```

	company	review
0	Accenture	[environment, peers]
1	Accenture	[environment, young, professionals]
2	Accenture	[access, big, projects, clients]
3	Accenture	[field, consultancy, company, start, career, a...
4	Accenture	[exposure, lot, projects]



```
[12]: import gensim.corpora as corpora

[51]: # Create Dictionary
      id2word_alldata = corpora.Dictionary(datalda_bigrams_trigrams)

      # Create Corpus
      texts_alldata = datalda_bigrams_trigrams

      # Term Document Frequency
      alldatalda_corpus = [id2word_alldata.doc2bow(text) for text in texts_alldata]

      # print the Document-Term Matrix
      alldatalda_corpus

[51]: [[(0, 1), (1, 1)],
       [(0, 1), (2, 1), (3, 1)],
       [(4, 1), (5, 1), (6, 1), (7, 1)],
       [(8, 1),
        (9, 1),
        (10, 1),
        (11, 1),
        (12, 1),
        (13, 1),
        (14, 1),
        (15, 1),
        (16, 1)
```

For example, it is observed that the first term of the first review, which is "environment" (0), is also repeated in the second review. For this reason, this term appears with frequency "1" in the first two documents within the DTM.

7.4.2.2 LDA model

It is observed that the method that implements the LDA model presents certain parameters to define how to train and generate the topic model. The main parameters are:

- Number of topics (k): main and important parameter to be able to define in some way.
- Alpha and beta: they are hyperparameters that affect the density of topics. The default value for both is 1/num_topics. With a higher alpha, documents are assumed to be made up of more topics and result in more specific topic distribution per document. with high beta, topics are assumed to made of up most of the words and result in a more specific word distribution per topic.
- Chuncksize: is the number of documents to be used in each training pass.
- Update_every: how often the model parameters are updated.
- Passes: The number of passes through the corpus during training.

I have run an initial model with 10 topics.

7.4.2.3 Model validation (metrics)

The metric used to validate the model is the ***"Coherence Score".*** The topic coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. It helps to distinguish between topics that are semantically interpretable and topics that are artifacts of statistical inference. Within Coherence metrics, I will use **"c_v"** measure, which is based on one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information and the cosine similarity.

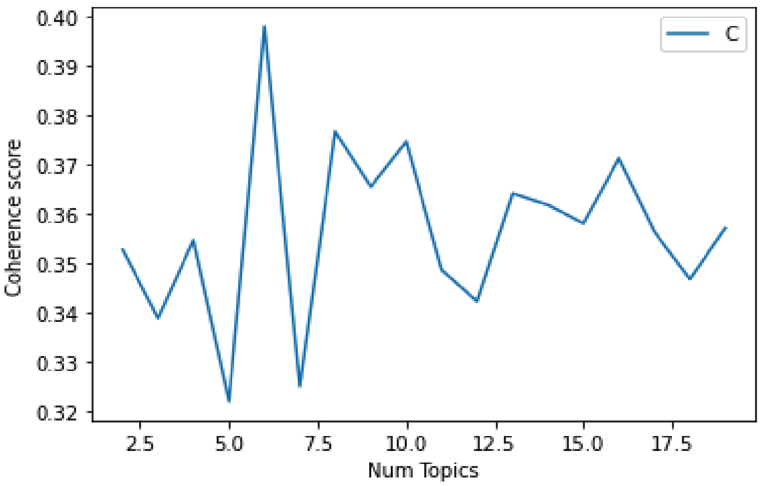
But the following question has arisen: how can I know how to improve the metric?

Since the LDA model requires that the number of topics to be searched be defined in advance, I consider it important to analyze this parameter before running the model. Therefore, I have based the optimization of the model on the parameter *k* or "number of topics"

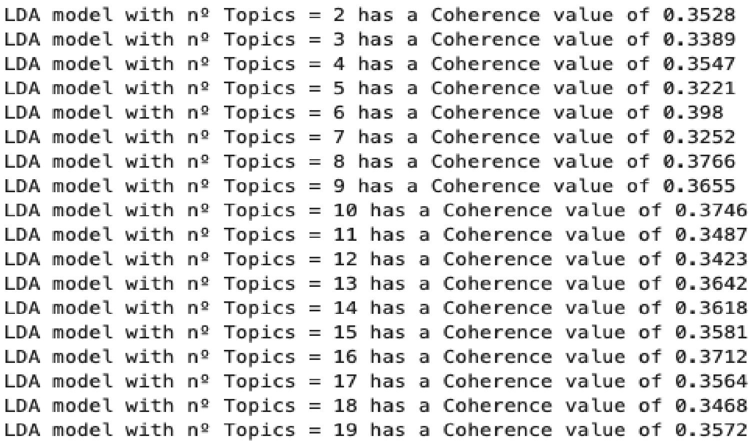
In this sense, the following question has arisen: what number of topics is the most appropriate to optimize the model?

I have been able to resolve this issue with what is known as ***"hyperparameter scoring"*** or *"parameter tuning".*

The following graph represents the different coherence values obtained for different *k* with which the LDA model was trained, up to a maximum of 20 topics (previously indicated):



The previous graph was obtained from the following values:



```
LDA model with nº Topics = 2 has a Coherence value of 0.3528
LDA model with nº Topics = 3 has a Coherence value of 0.3389
LDA model with nº Topics = 4 has a Coherence value of 0.3547
LDA model with nº Topics = 5 has a Coherence value of 0.3221
LDA model with nº Topics = 6 has a Coherence value of 0.398
LDA model with nº Topics = 7 has a Coherence value of 0.3252
LDA model with nº Topics = 8 has a Coherence value of 0.3766
LDA model with nº Topics = 9 has a Coherence value of 0.3655
LDA model with nº Topics = 10 has a Coherence value of 0.3746
LDA model with nº Topics = 11 has a Coherence value of 0.3487
LDA model with nº Topics = 12 has a Coherence value of 0.3423
LDA model with nº Topics = 13 has a Coherence value of 0.3642
LDA model with nº Topics = 14 has a Coherence value of 0.3618
LDA model with nº Topics = 15 has a Coherence value of 0.3581
LDA model with nº Topics = 16 has a Coherence value of 0.3712
LDA model with nº Topics = 17 has a Coherence value of 0.3564
LDA model with nº Topics = 18 has a Coherence value of 0.3468
LDA model with nº Topics = 19 has a Coherence value of 0.3572
```

This metric will evaluate the coherence score of each model for a maximum number of topics previously indicated. So you can compare how this metric changes as the number of topics in the model increases, up to the maximum number of topics indicated. And then select the appropriate number of topics for our LDA model, but which will be? the one with the highest coherence score (c_v)?.

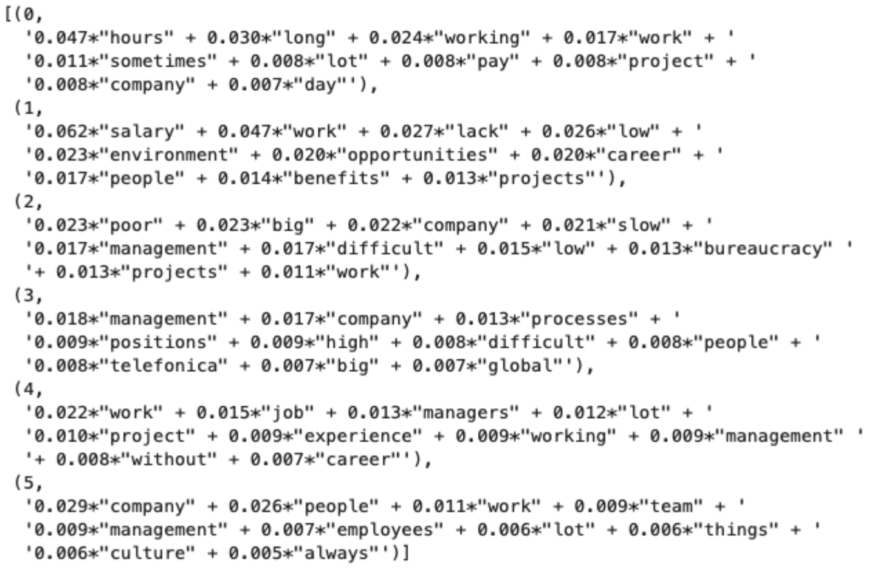
7.4.2.4 Model fitted

The optimized model will be the one with the number of topics that has a higher coherence score *(c_v)* according to the hyperparameter score, but considering also the context. That means, the optimized model will not only be the one with the number of topics that has the highest consistency score *(c_v)* based on the hyperparameter score.

The coherence value is a simple way to see how good the model is but, for the choice of the optimal *k* value, the *c_v* does not have to be simply the largest, but also the amount of data that is being processed is important. Regarding this, if we see the same keywords repeated in many of the topics, it is probably a sign that the parameter *k* is too large.

For this reason, it is best to choose the parameter *k* that gives the maximum coherence value before the curve begins to flatten out.

So I have executed a fitted model, indicating the optimal number of topics. The LDA model was built with **6** different topics, where each of these topics is a combination of keywords and each keyword contributes a certain weight to the topic. We can see each keyword of each topic and the weight or importance of each keyword as shown below:



```
[(0,
  '0.047*"hours" + 0.030*"long" + 0.024*"working" + 0.017*"work" + '
  '0.011*"sometimes" + 0.008*"lot" + 0.008*"pay" + 0.008*"project" + '
  '0.008*"company" + 0.007*"day"'),
 (1,
  '0.062*"salary" + 0.047*"work" + 0.027*"lack" + 0.026*"low" + '
  '0.023*"environment" + 0.020*"opportunities" + 0.020*"career" + '
  '0.017*"people" + 0.014*"benefits" + 0.013*"projects"'),
 (2,
  '0.023*"poor" + 0.023*"big" + 0.022*"company" + 0.021*"slow" + '
  '0.017*"management" + 0.017*"difficult" + 0.015*"low" + 0.013*"bureaucracy" '
  '+ 0.013*"projects" + 0.011*"work"'),
 (3,
  '0.018*"management" + 0.017*"company" + 0.013*"processes" + '
  '0.009*"positions" + 0.009*"high" + 0.008*"difficult" + 0.008*"people" + '
  '0.008*"telefonica" + 0.007*"big" + 0.007*"global"'),
 (4,
  '0.022*"work" + 0.015*"job" + 0.013*"managers" + 0.012*"lot" + '
  '0.010*"project" + 0.009*"experience" + 0.009*"working" + 0.009*"management" '
  '+ 0.008*"without" + 0.007*"career"'),
 (5,
  '0.029*"company" + 0.026*"people" + 0.011*"work" + 0.009*"team" + '
  '0.009*"management" + 0.007*"employees" + 0.006*"lot" + 0.006*"things" + '
  '0.006*"culture" + 0.005*"always"')]
```

I can observe that 6 main topics or themes are discussed in the reviews, which I could name as follows:

- <u>Topic 1</u>: **"schedule"** (it is observed that long working hours are mentioned a lot)

- Topic 2: **"remuneration package and career plan"** (salary, benefits, professional development opportunities, etc.)

- Topic 3: **"bureaucracy"** (it is observed that high bureaucracy is mentioned, which implies slowness in decision making and in the execution of strategic plans)

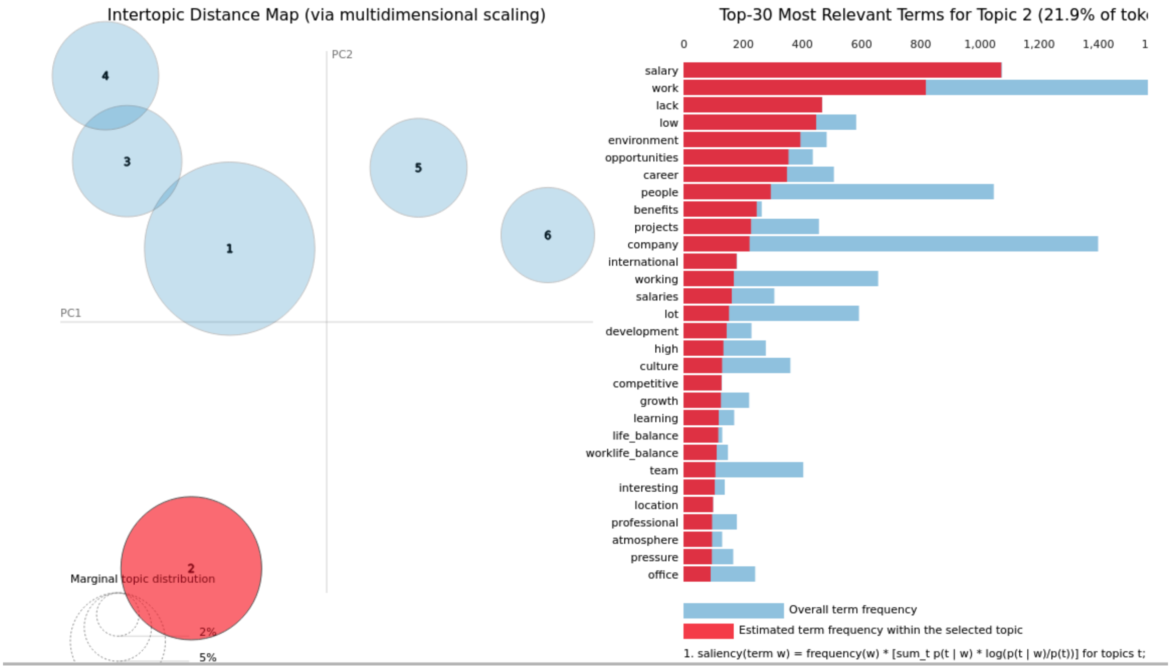
- Topic 4: **"internal processes"**

- Topic 5: **"management team"** (talking about bosses)

- Topic 6: **"culture and people"**

7.4.2.5 Topics visualization

To display the topics mentioned and their keywords, I have used the **pyLDAvis** library. The plots are in the notebook itself as code return, and it is easy to interact with. The mode of interaction is explained in the following section about the user manual (7.4.4).



I can observe that **6 main topics** or themes are discussed in the reviews, which I could name as follows:

- Topic 1: **"culture and people"**

- Topic 2: **"remuneration package and career plan"** (salary, benefits, professional development opportunities, etc.)

- Topic 3: **"management team"** (talking about bosses)

- Topic 4: **"schedule"** (it is observed that long working hours are mentioned a lot)

- Topic 5: **"internal processes and politics"**

- Topic 6: **"bureaucracy"** (it is observed that high bureaucracy is mentioned, which implies slowness in decision making and in the execution of strategic plans)

In the previous visualization of the fitted LDA model on pyLDAvis, topic 2 has been selected which, as I mentioned, refers to the remuneration package, career development and also to the work-family balance as can be seen, since tokens such as "worklife_balance" and "life_balance" are very representative in this topic.

Also it is observed that the bubbles are large, which indicates a great representativeness of the topics in the corpus. We also see that they are distributed throughout the quadrants and only those corresponding to topics 1, 3 and 4 overlap slightly, which is due to the fact that there are words that share high representativeness in these three topics, such as: "work", "company " and “management". But it is not relevant.

The interaction with the dashboard is explained in the section about user manual (7.4.4.).

7.4.3 ANALYSIS BY COMPANY

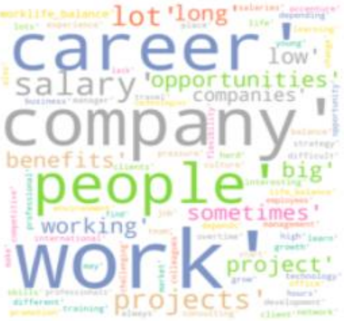
In this section, I will not explain the steps for creating the topic model or the code, since it has already been explained in the previous section. In this I will focus on making a comparison between companies and the conclusions obtained from it.

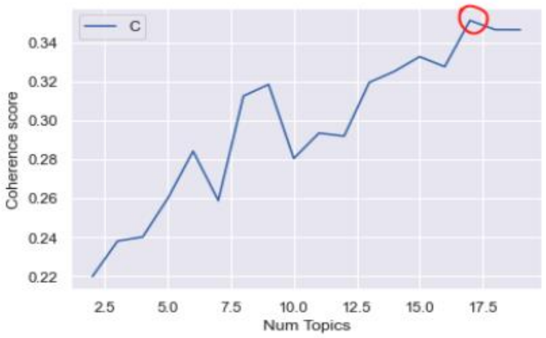
As I mentioned in the section 4, I will compare three groups of companies or industries, in relation to my purpose:

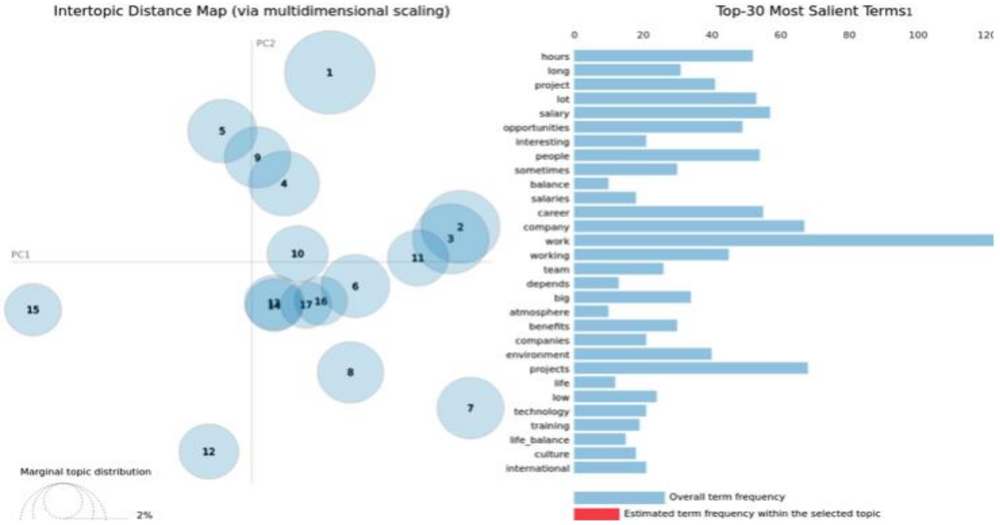
- IT Consunting
- Business consulting
- Own IT product (startups unicorns based in Barcelona)

IT CONSULTING

- ACCENTURE







In the counting of words it is observed that, beyond the usual words that are the most representative in general (job, company, people, etc.), it can be seen how words related to what the company offers are highly named: salary, benefits, career opportunities and projects, etc.

On the other hand, it's curious how people are named more than salary.

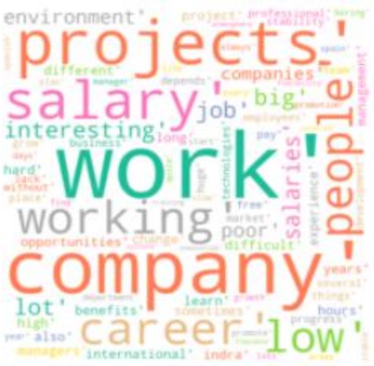
I have created a fitted model with a *k* value of 17 (number of topics), with a coherence value of 0.3514 (the lowest of the four IT consulting companies compared), following what was observed in the graph of the hyperparameter scores.

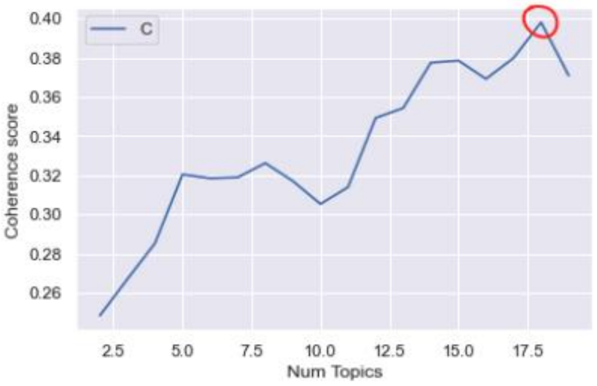
However, despite the fact that this is the most appropriate number of *k* according to this metric, in the visualization we see how several topics overlap almost completely on the right side of the graph, which indicates that a large part of the representative words of these topics are shared. It happens between topics 13 and 14 that commonly talk about career opportunities, although 14 focuses on the balance or conciliation between personal and work life; as well as between topics 16 and 17 since they share relevant words, but it is seen how 17 refers to working hours and holidays, while 16 refers to international projects and environments.

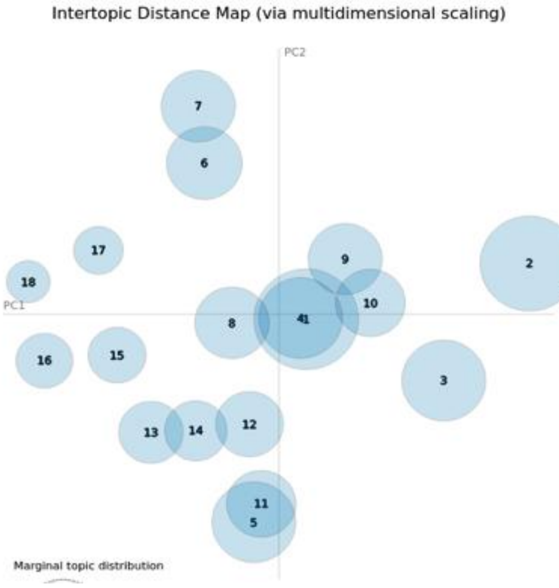
At a general level, it is significant how the words "hours" and "long" are widely mentioned, while reference is also made to "life balance". This may indicate that employees are complaining about long working hours and overtime.

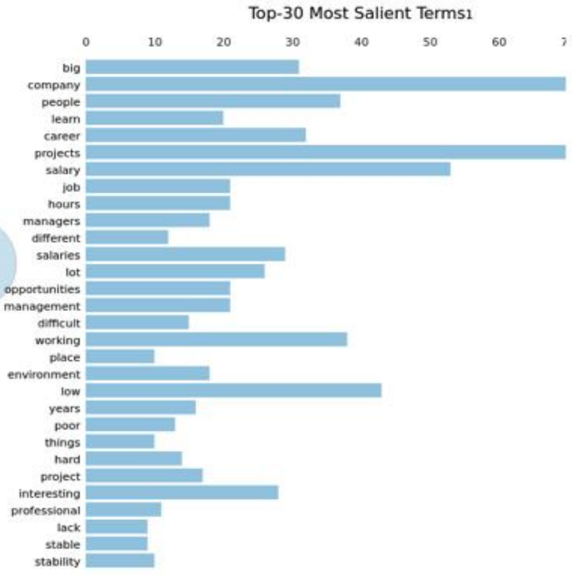
- INDRA / MINSAIT

The reviews of both companies have been grouped in the same dataframe, to carry out the joint analysis, since they belong to the same business group.









It seems that there is a high reference to interesting projects, and salary is also mentioned a lot (more than in Accenture, in which people is more relevant).

I have created a fitted model with a *k* value of 18 (number of topics), with a coherence value of 0.398, following what was observed in the graph of the hyperparameter scores.

Regarding the visualization of the model, at a general level, it can be highlighted how much is discussed about management, referring to bosses or directors, and about stability.

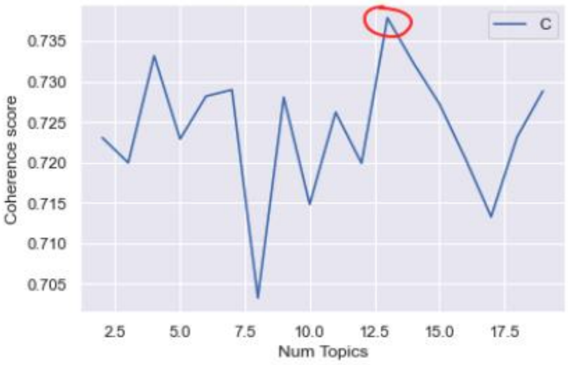
It is observed how the bubbles are more spread out over the four quadrants and overlap less than in the case of Accenture, which is positive because it indicates that the topics have less words shared between them and, therefore, they refer to more independent themes.

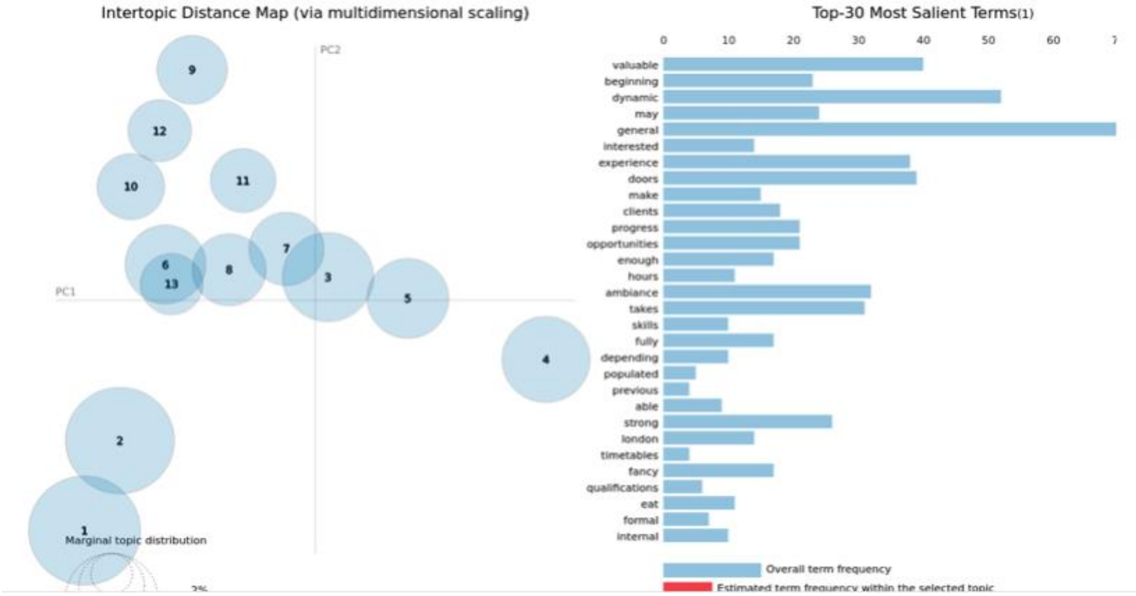
The topics that overlap the most are 1 and 4, since they share relevant and generalized words such as "work" and "projects", but it can be seen how 1 refers to the variety of projects and clients; while 4 refers to projects management and promotion opportunities within them. On the other hand, although to a lesser extent, topic 10 also overlaps, which refers to the stability of the projects and security in the company, and also topic 9, which could refer to the technological environment of the projects.

Topics 5 and 11 also overlap, since they both talk about training.

- EVERIS

salary
company
people
work
career
projects
environment
lot
learn
team
working
sometimes
companies
management
experience
hours
low





It is observed how people and salary appear with the same relevance, and both with greater representativeness than in Accenture and Indra. And it seems that there is a lot of talk about the environment, learning and career.

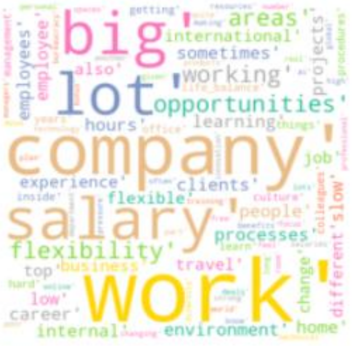
I have created a fitted model with a *k* value of 13 (number of topics), with a coherence value of 0.7379 (the highest coherence value at the moment), following what was observed in the graph of the hyperparameter score.

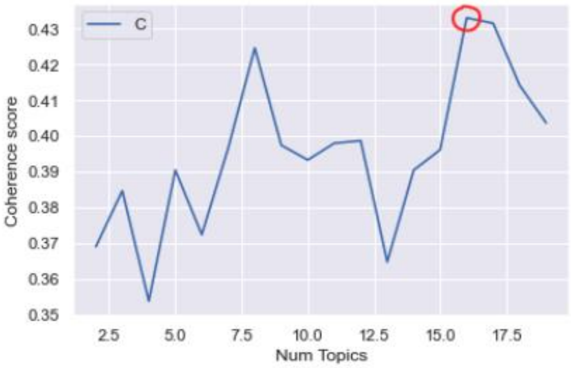
It is observed that the 13 topics are properly distributed across the quadrants and that they barely overlap, with the exception of topics 6 and 13. It seems that both refer to dynamism and internal opportunities for progress and to obtain experience and skills in other areas, which

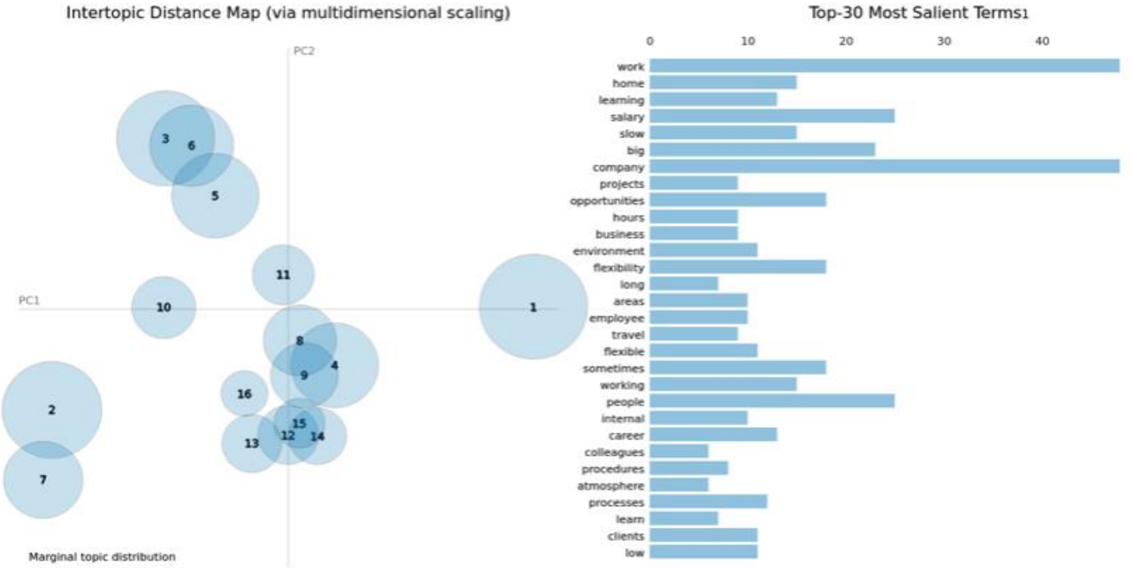
which can be related to more liquid, agile or holocratic organizational structures. The difference between the two is that 6 refers to new skills and 13 refers to other environments and clients.

They are not the only topics that speak of opportunities and progress, since the three great topics that are far from the majority in the quadrant (1, 2 and 4), also speak of growth. 1 and 2 are closer to the acquisition of new projects and partners, at the business level; while 4 seems to refer more to employees, to performance evaluations.

- <u>IBM</u>







Apart from the usual words, it can be seen how in this case the opportunities, flexibility and international environment stand out.

I have created a fitted model with a *k* value of 16 (number of topics), with a coherence value of 0.43, following what was observed in the graph of the hyperparameter score.

Having a first look, small bubbles can be seen that overlap almost touching the center of the quadrants, corresponding to less representative topics. So I think the model could be improved.

In general aspects, it can be seen how the company is in which flexibility is mostly talked about and the word "home" is also representative, so we can assume that flexibility for remote work is talked about in a large part. Topic 4 mainly refers to this.

On the other hand, we also see that it is the company in which internal procedures and processes are most talked about, surely referring to bureaucratic aspects. Topic 8 refers to this issue and the slowness as a consequence.

Topics 7 and 3 speak mainly about career and training plans, the latter also about recognition.

Topic 2, which is quite representative, refers to business, new contracts and a high number of subcontractors.

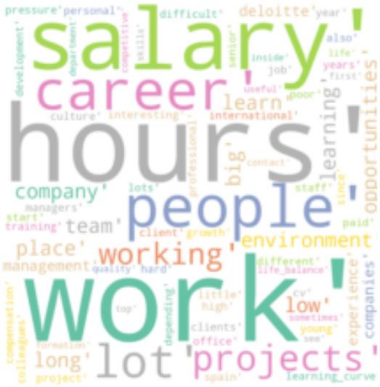
The isolated topic 10 talks for example about locations, infrastructure and facilities. Topic 11, which is also isolated, seems to refer to the loss of projects in recent years. Although both have less representativeness over the total.

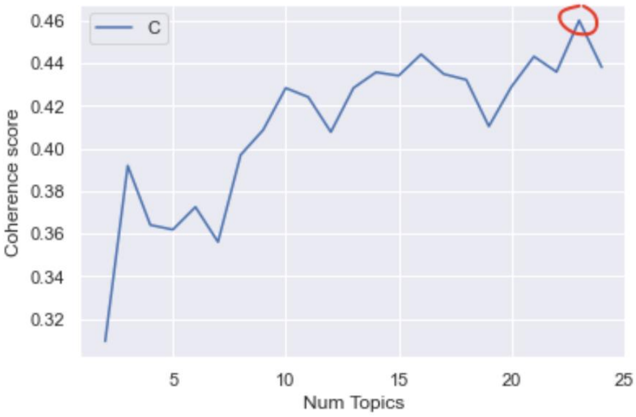
Conclusions: It can be seen that about this industry, employees talk about variety of projects, opportunities and flexibility as positive aspects; as well as about salaries, management and bureaucracy on the negative side.

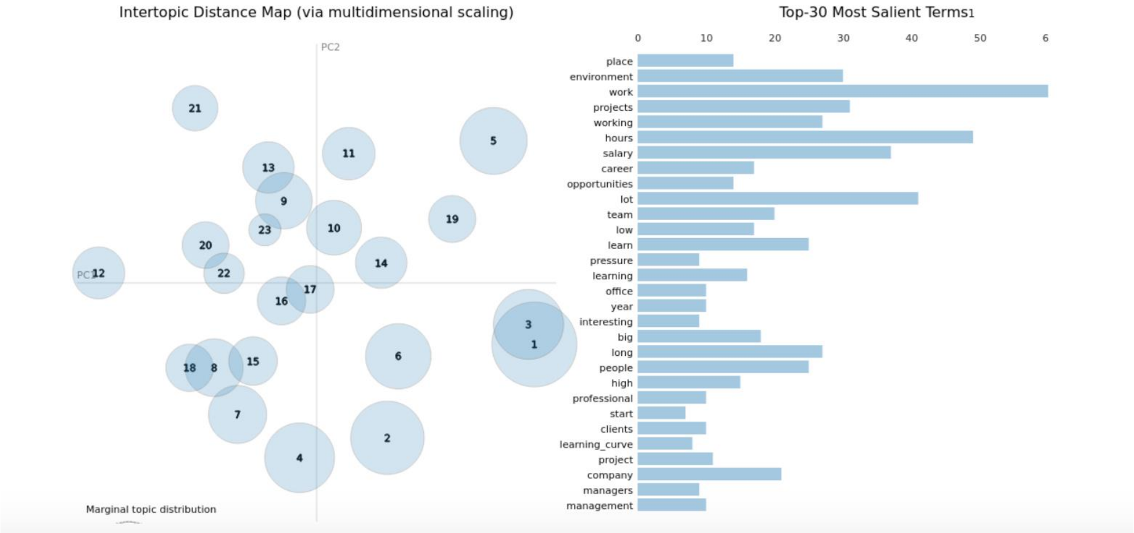
Among the companies analyzed, it is observed that about Everis they talk more about the environment, people, training and career plan. There is talk of high flexibility about IBM, as well as high bureaucracy and slow procedures. About Accenture, high working hours are mentioned. And about Indra they talk about the variety of projects, the stability in the company and about the managers as a negative side.

BUSINESS COMSULTING

- DELOITTE





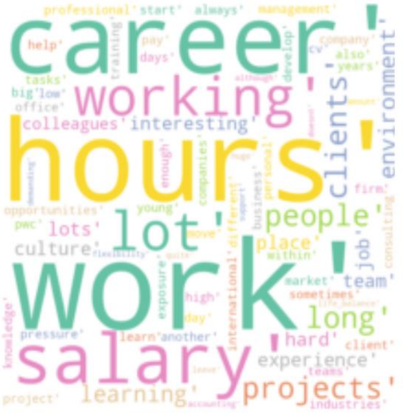


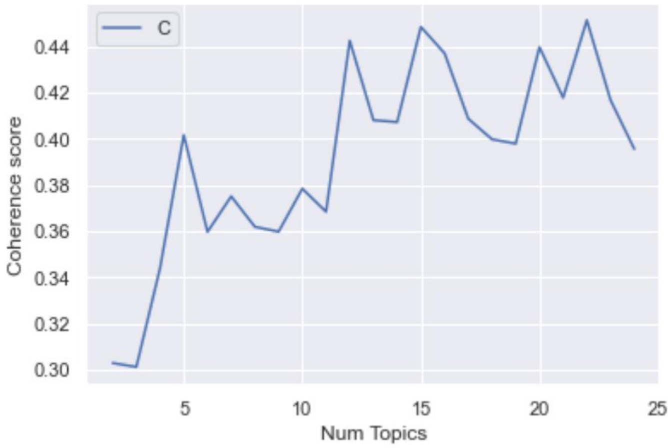
I have created a fitted model with a *k* value of 23 (number of topics), with a coherence value of 0.46, following what was observed in the graph of the hyperparameter score.

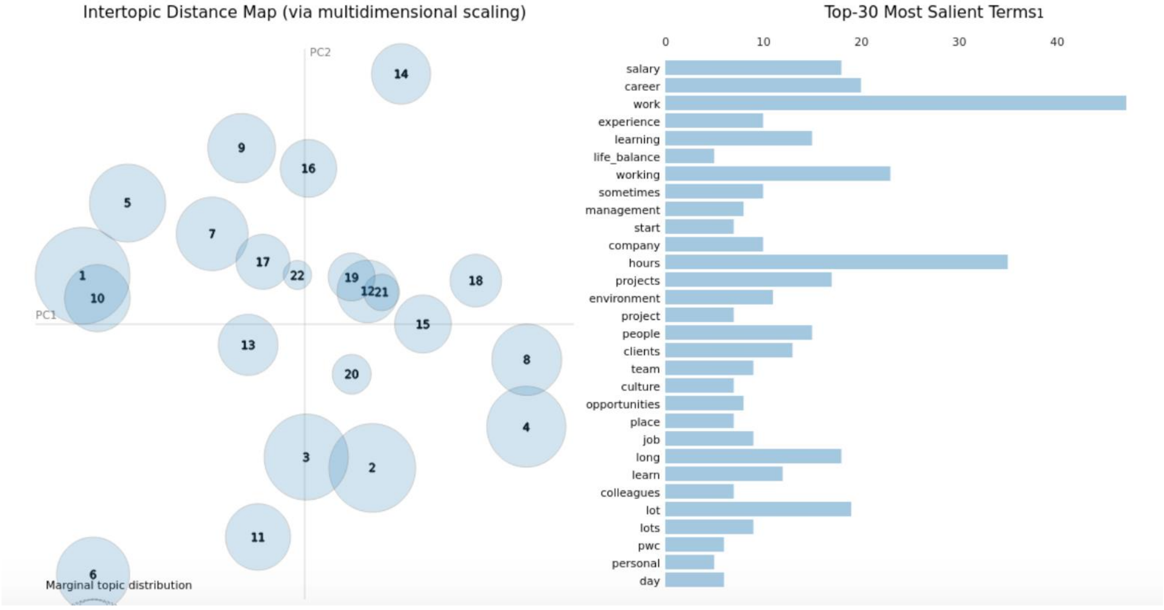
The fact that the word "hours" is one of the most relevant, as we can see at a glance in the "word cloud", predicts the topics that we are going to find.

In this case, it is observed that there are many topics with low representativeness, which is reflected in the existence of so many small bubbles. The curious thing in this case is that it does not seem that the topics overlap, except for certain cases, the most prominent being with topics 1 and 3, as they are relevant. In both cases, are talking about long working hours, in case 1 it is more related to poor work-life balance and in case 3 to salary. Topics 2, 5 and 6, which are also highly relevant, also speak of long working hours. And another less relevant, such as 14 and 8, also talk about overtime, the first referring again to balance with personal life and the second to projects.

- PWC







I have created a fitted model with a *k* value of 22 (number of topics), with a coherence value of 0.45, following what was observed in the graph of the hyperparameter score.

It happens as in the previous case with Deloitte. Many and small bubbles are observed, so many topics have low representativeness over the total. The word "hours" is again the most representative behind the more generalized ones ("work", "company", etc.), and is even repeated more than salary.

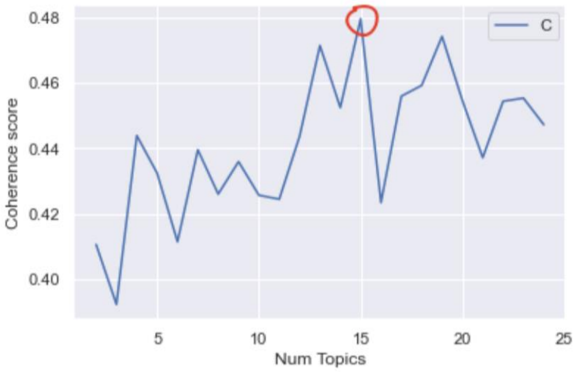
That is, talking again about long hours and long working days in some of the most relevant topics (1, 5 and 6).

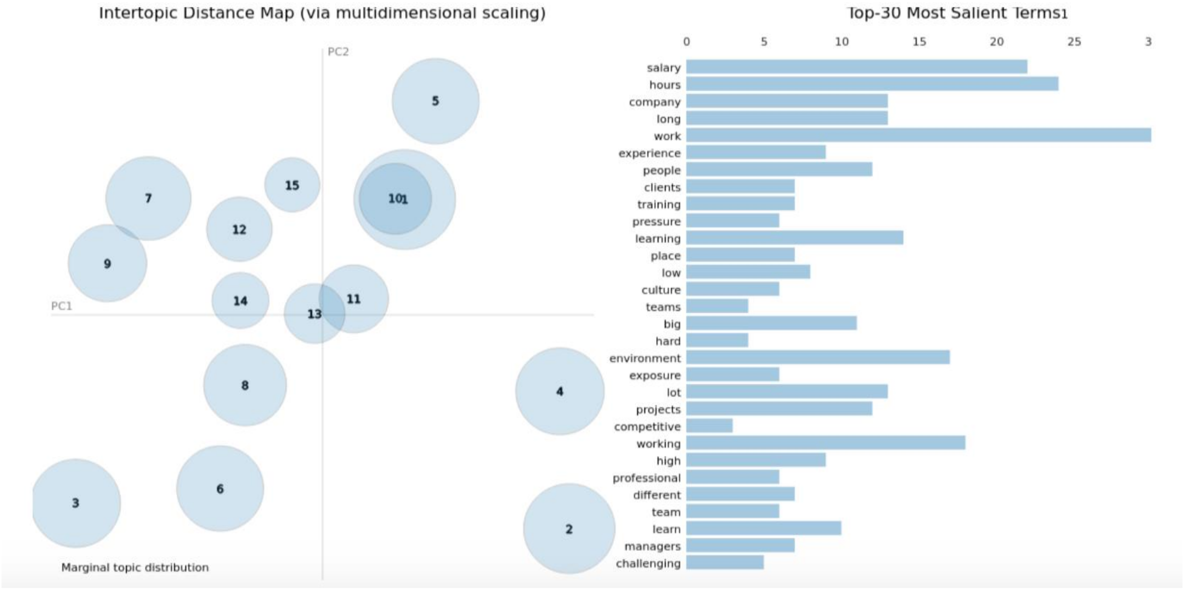
On the other hand, the topic 2 refers to "hard work"; and the 7 to "extra hours", high pressure and lack of conciliation (that is understood). Topic 10 also talks about pressure, and 11 talks about high demand and stress.

Regarding to Deloitte, PWC talks more about training and learning, in relevant topics such as 4 and 8. And career opportunities in 16.

- <u>KPMG</u>

salary
hours
working
learning
work
environment
company
management
learn
experience
team
different
international
long
job
people
high
colleagues
training
lot
culture
clients
low
partners
projects
opportunities
place
pressure
big
days
young
junior
managers
challenging
demanding
career
firm
business
skills
life_balance





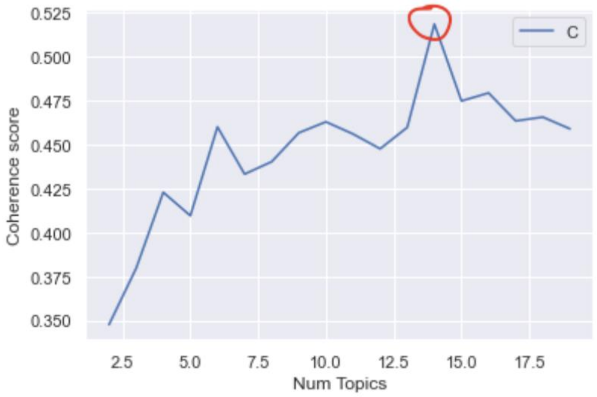
I have created a fitted model with a *k* value of 15 (number of topics), with a coherence value of 0.4795, following what was observed in the graph of the hyperparameter score.

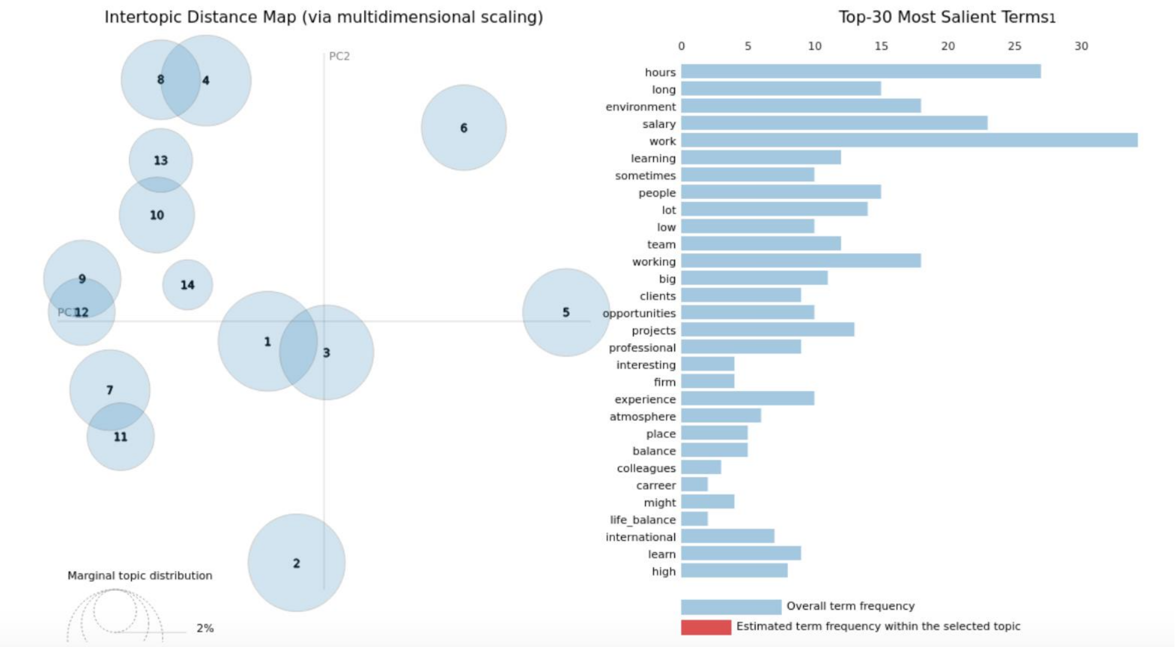
Long working hours and long days are spoken of again, in relevant topics such as 8 and 7, followed by 11, 13 and 14. It can be seen how most overlap in the center of the quadrant.

On the other hand, topics 7 and 12 talk about pressure and a competitive environment. The 10 also speaks of pressure but also of high exposure (to business and clients, I mean). The 5 speaks of a young environment. The 4 speaks of business, projects and partners. The 1 speaks of training. And the topic 6 refers to a professional and challenging environment, with international projection and salary compensation.

- E&Y

learn
experience
willing
atmosphere
companies
salary
people
development
employee
low
high
different
career
firm
environment
projects
international
lot
work
compared
job
balance
sometimes
big
hours
clients
learning
opportunities
pressure
company
grow
year
salaries
professional
long
team
place
client
working
management





I have created a fitted model with a *k* value of 14 (number of topics), with a coherence value of 0.5185 (the highest of the compared business consulting companies), following what was observed in the graph of the hyperparameter score.

Once again, talking about many hours of work, as can be seen mainly in topics 5 and 6, which are quite relevant. In 5 we also see that it is related to two new bigrams: "tight_deadlines" and "busy_season", so it is understood that there is a lot of work and pressure to speed up deliveries. The 6 speaks of pressure, again.

Topic 4 talks about people, team, and opportunities. The 2 refers to low salaries. The 12 refers to the work environment and, once again, competitiveness is named.

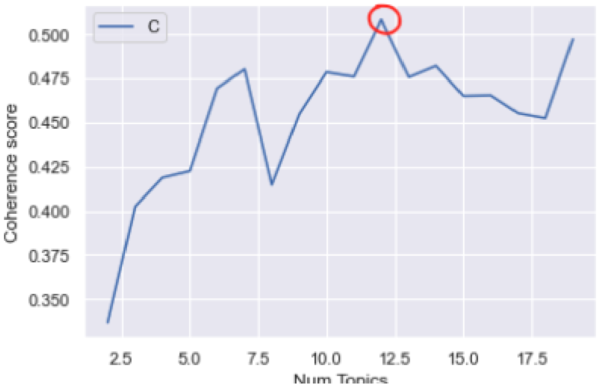
Conclusion: If something is clear to us from this analysis, it is that in the business consulting industry, long hours are worked, there is high pressure, tight deadlines, low conciliation and a competitive environment; while on the other hand the professional, challenging environment and business and networking opportunities are valued. And that is shared by all companies.

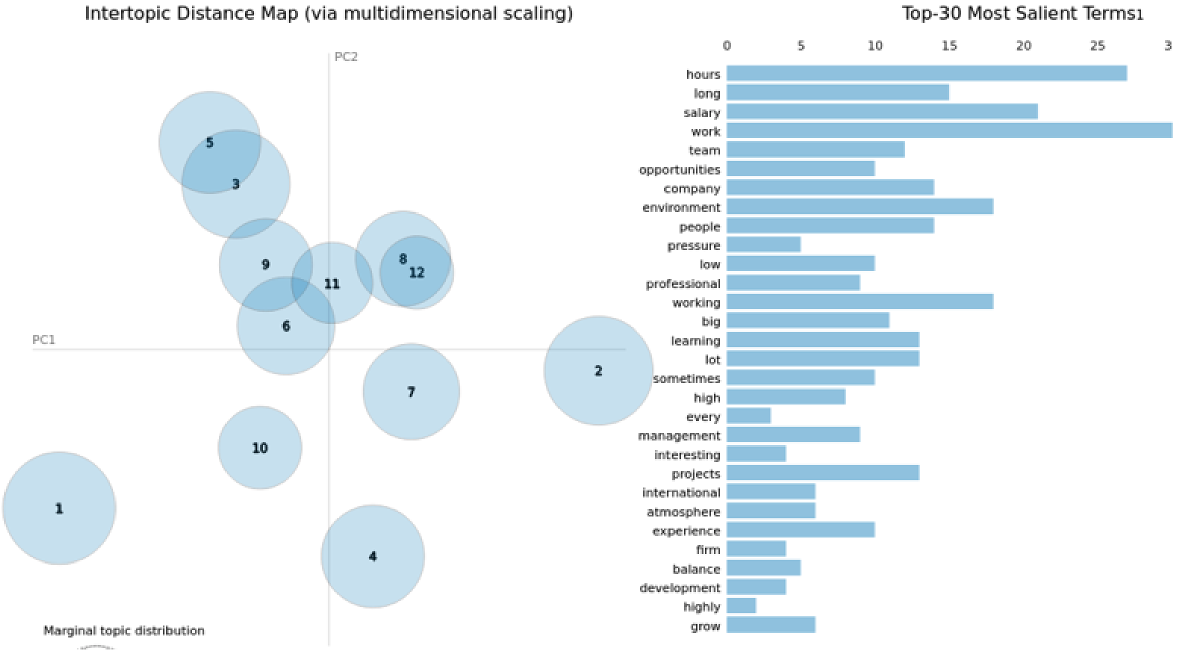
Within these companies, it is perhaps at PWC where career and training opportunities are most talked about.

OWN IT PRODUCT STARTUPS

- GLOVO

opportunities
balance
development
clients
people
team
international
willing
learn
atmosphere
work
hours
projects
sometimes
salaries
client
company
low
challenging
year
learning
long
big
place
career
environment
companies
pressure
professional
salary
experience
compared
job
grow





I have created a fitted model with a *k* value of 12 (number of topics), with a coherence value of 0.5082, following what was observed in the graph of the hyperparameter score.

In the case of Glovo, long working hours are also mentioned above the rest of the terms, as we can see mainly in topics 1 and 4, in which overtime is also mentioned. Low salaries are also mentioned to a large extent, as can be seen in topics 2 (where pressure and competitiveness are also discussed) and in topic 7 (in which conciliation or work-life balance is also mentioned).

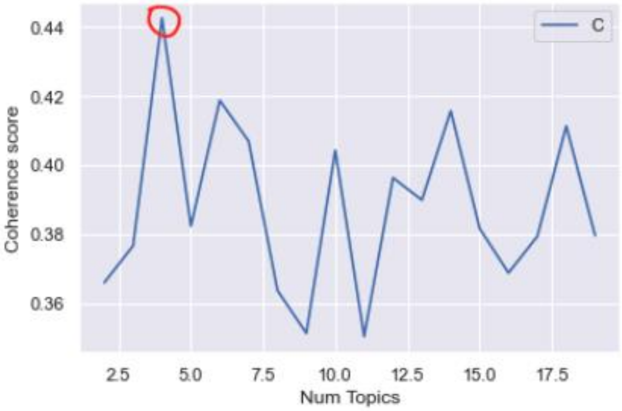
Pressure is also mentioned in topic 11, which also talks about a pyramidal structure and high rotation.

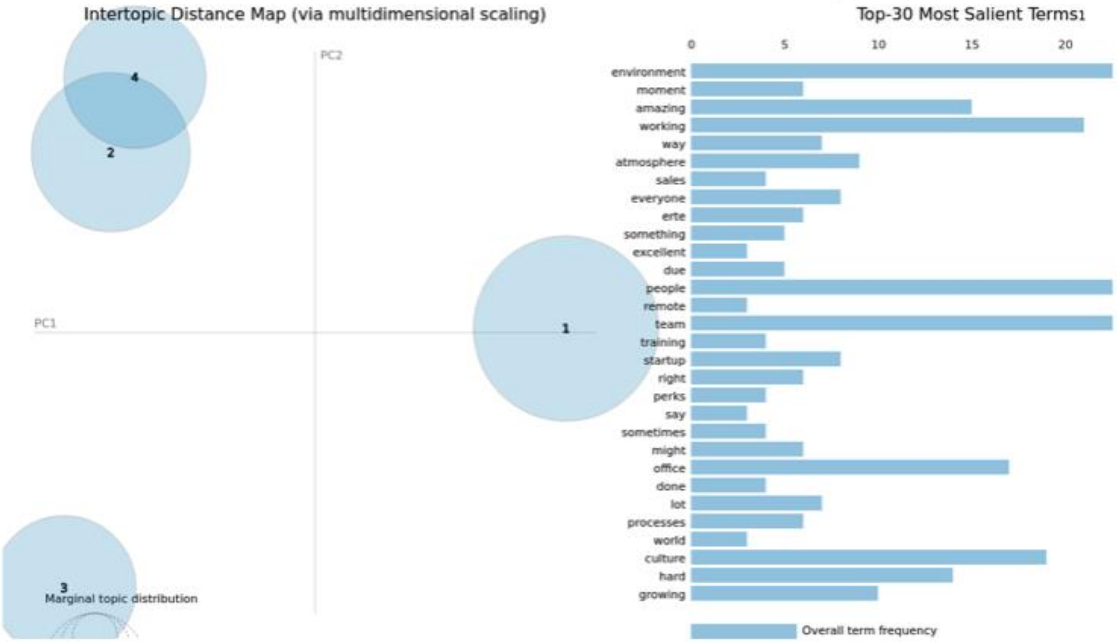
On the other hand, topic 8 refers to a young environment of colleagues. Topics 6 and 9 talk about professional development.

In this case, I consider it to be a "good" model, the bubbles are distributed throughout the quadrants, they are representative and, although some overlap to a certain extent, the topics can be easily distinguished.

- <u>TRAVELPERK</u>

culture
company
people
growing
environment
team
work
working
office
colleagues
everyone
grow
leadership
friendly
startup
place
right
management
product
experience
amazing
employees
hard
travelperk
teams
travel
atmosphere
learn
always





I have created a fitted model with a *k* value of 4 (number of topics), with a coherence value of 0.4427, following what was observed in the graph of the hyperparameter score.

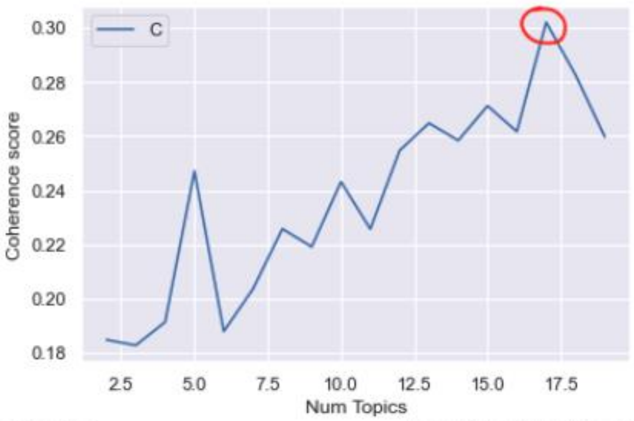
In a first look at the 'word cloud', we see that the words "culture", "growth" and "environment", unlike the majority of companies analyzed before, will be very named in the topics, above the rest.

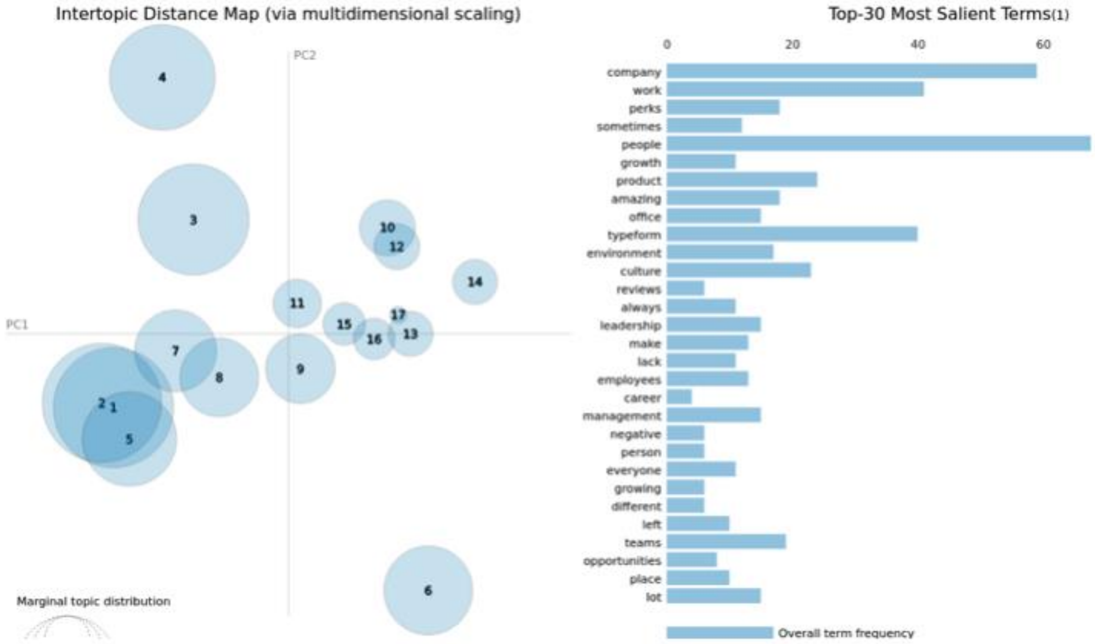
In the case of Travelperk, it is the first time that we see a model with such a low number of topics. We can differentiate four major topics, with approximately the same representativeness within the corpus (topic 1 perhaps has a bit more), and that differ at a semantic level except for topics 2 and 4. Therefore, both talk about people and environment. Topic 4 makes more reference to the environment in the workplace or offices, where teamwork, friendliness, colleague is breathed; while growth, flexibility and remote work are mentioned. While topic 2 talks more on a cultural and company level, since it refers to transparency and leadership.

You have the feeling that the topics could be delimited to a greater extent, although it is true that they all talk about people and culture, as well as product and growth; and this is the most important point. Since it is understood that they are the great attractions of this company for its employees.

- <u>TYPEFORM</u>

typeform
product
company
working
also
things
perks
teams
work
amazing
people
culture
team
leadership
lot
office
environment
hard
everyone
know
growth
employees
experience
sometimes
colleagues
companies





I have created a fitted model with a *k* value of 17 (number of topics), with a coherence value of 0.302 (the lowest of all the models created), following what was observed in the graph of the hyperparameter score.

In the case of Typeform, something similar to the previous case happens, because the culture and above all the people and the team are mostly named. In this case, the product also stands out.

We see that we have in this case 17 topics distributed mainly in the center of the quadrant, 8 of these with less representativeness that are semantically close, most of which speak about the product and growth.

On the other hand, on the opposite side, it is noted that there are three topics of high representativeness that overlap each other: 1, 2 and 5. That means, all three talks about people, team and culture. But it is understood that 5 refers more to leadership and talent of the engineers, 1 to new product ideas, and 2 to caring for teams and their growth.

Conclusion: In the comparison, while Typeform and TravelPerk are valued positively, the first above all for its product and the talent of its teams, and the second for its culture, good environment and growth opportunities; Glovo is very poorly valued (overtime, high turnover, low conciliation, high pressure, competitiveness, large pyramidal structure, etc.).

Globally, startups develop their own product and in the end they offer that something different (there is talk of business ideas, teams of powerful engineers, etc). They are more dynamic, more innovative, there is less bureaucracy and less hierarchical structure because it has yet to be defined, which translates into more opportunities for growth. In short, this mix is more attractive in the market.

7.4.4 VISUALIZATION USER MANUAL

The pyLDAvis visualization is designed to help interactively with:

- Better understanding and interpreting individual topics. You can manually select each topic to view its top most frequent and/or “relevant” terms, using different values of the λ parameter. This can help when you’re trying to assign a human interpretable name or “meaning” to each topic.
- Better understanding the relationships between the topics. Exploring the intertopic Distance Plot can help you learn about how topics relate to each other, including potential higher-level structure between groups of topics.

The key aspects to understand and interact with the visualization of a topic model in pyLDAvis are the following:

- Each bubble on the left represents a topic.
- The closeness between bubbles reflects the semantic closeness between the topics.
- Axis do not give us any information.
- The bigger the bubble, the more predominant that topic is. That is, the higher percentage of the number of reviews in the corpus is about that topic.
- A good topic model is one that has fairly large, non-overlapping bubbles scattered throughout the graph rather than being all together and clustered in a single quadrant.
- A model with many topics will surely have small bubbles, located in the same region of the graph and with many overlapping cases.
- If we move the cursor over each of the bubbles, the words and bars in the graph on the right will update. These words are the most important keywords that make up the selected topic.
- Blue bars represent the overall frequency of each word in the corpus. If no topic is selected, the blue bars of the most frequently used words will be displayed.
- Red bars give the estimated number of times a given term was generated by a given topic.
- If we move the cursor over each of the words (from the bar chart on the right), we can see how the bubbles of those topics where these words have a greater representativeness automatically enlarge.

7.4.5 OBSERVATIONS

In this section I want to evaluate the topic models created, add some reflections.

To evaluate the model, we must ask ourselves the following questions:

- Has it captured the internal structure of the corpus?
- Are the topics understandable?
- Are the topics consistent or coherent?
- Does it serve the purpose it is being used for? (if the use case defines a "good" model)

I think that the global model is a "good" model, the most representative words and topics are consistent with respect to the corpus data. And you can properly understand the differences between the 6 topics. On the other hand, of course, it helps in the purpose of this project.

Regarding the models by company, however, I consider that they can be adjusted to a greater extent, especially those that have many topics with less representativeness and whose terms overlap (because they share semantics). But, for this, I consider it necessary to have a greater volume of reviews per company. Anyway, these models have also been very useful for the project purpose.

We have to consider the difficulty in evaluating a topic model (a qualitative evaluation is tedious and subjective, and quantitative ones do not always produce understandable results or words sometimes do not capture the semantics of a category either); and that there is no universally accepted "good" model, but it depends on the purpose. In this case they have been useful.

7. FRONTEND WITH TABLEAU

In addition to plotting visualizations throughout the project to understand the data in the notebooks themselves, using different Python libraries; I have also made dashboards in Tableau to visualize information in cases in which I have considered that this tool could provide me with a more agile, intuitive and interactive visualization.

Regarding this, I have used Tableau to visualize the information specified below:

A. After cleaning the main dataset obtained in the scraping process, apart from the python plots that can be seen in the notebook itself, I have used Tableau with the aim of performing an **initial exploratory analysis (6)** in a faster, visual and interactive way.

   - Dataset: ***'companies_reviews_v3.csv'***
   - Total reviews: 9,989 (in spanish and English)
   - Link to dashboard:

     https://public.tableau.com/shared/X3JPNQJP3?:display_count=n&:origin=viz_share_link
   - Title: EDA 1 (from clean initial dataset)

   The interesting thing is that you can interact with the graphs together, filtering if you select, for example, by year, by language, by company, by region or by location.

   It can be seen that the dashboard is divided into five quadrants or graphs:

- Counts of reviews - evolution by year: we can observe the evolution of the number of reviews made since Glassdoor was released in 2008 until the first quarter of 2021 (when I finished the scraping process). An upward trend can be observed, especially from 2018. Obviously, the drop in 2021 is only due to the fact that there are only reviews from the first quarter of that year in the dataset.

- Counts of reviews by language: of a total of 9.989 reviews, it can be seen that 6.298 (63%) are in Spanish and 3.691 (37%) are in English.

- Ranking top 15 companies - counts of reviews evolution by year: this is an interactive ranking, through which you can visually observe how the number of reviews made on Glassdoor evolves between 2008 and 2021, filtered by the 15 companies with the highest number of reviews and sorted in descending order based on the number of reviews.

  Let us remember that the dataset has been compiled with reviews of companies with high activity and relevance in Barcelona, in relation to the purpose of the project. It is observed that the companies that gather the largest number of reviews are coming from sectors or groups of companies that have been analyzed in the comparison: IT consulting (Accenture, Indra, Everis, IBM, etc.); business consulting (Deloitte, KPMG, PwC, etc.); plus some startups (Glovo, for example) and final companies (Amazon, for example) relevant in this location.

- Counts of reviews by company: in relation to what was mentioned in the previous point, this graph shows how effectively the companies with the most presence in the dataset by volume of reviews are Accenture, Deloitte, Everis, Indra, Amazon, etc., by that order.

- Counts of reviews by region and location: the total number of reviews collected by community or region and by province is reflected. It is observed that they are mainly concentrated in Madrid (5,168 / 52%), followed by Barcelona (3,925 / 39%). This means that the sum of reviews made in the rest of the regions gather only 9% of the total reviews in the dataset.

B. In **sentiment analysis (7.2.3),** to compare sentiment scores between companies.

- Dataset: ***'sentimentcounts_percompany.csv'***

- Total reviews: 7.381. Only in English, considering both features separately as different reviews. In other words, as explained in the corresponding section, there are a total of 3,691 reviews (rows) in English in the initial dataset, divided into two features ('pros' and 'cons'). In this dataset, both features have been considered as independent reviews and, therefore, they have been joined, resulting in a dataset with just twice as many reviews in English (7,381).

- Link to dashboard:

  https://public.tableau.com/views/EDA2SentimentAnalysis/Dashboard1?:language=es-ES&:display_count=n&:origin=viz_share_link

- Title: EDA 2 (Sentiment Analysis)

The dashboard is already very visual at a glance. But, to interact with it in a best way, a company can be selected (on the X axis of the bar chart) and, the proportion of scores for each sentiment over the total number of reviews about the selected company collected in the dataset, will automatically be filtered in the “pie chart” and in the “bar chart”.

We can see that, as we select a different company, these proportions change.

To make the observations, the volume of the type of sentiments over the total reviews between companies should not be compared, since each company has a different weight over the total reviews of the dataset. For this reason, if the comparison is made in this way, the most normal thing is that more reviews with positive sentiment are observed in "Accenture", for example, without considering that this is the company with the highest number of reviews in the dataset.

Therefore, the proportion of each type of sentiment over the total number of reviews of the company in question must be observed. And late, evaluate if the percentage of positive reviews over the total reviews in Accenture is higher or lower than in Indra, for example, because they are competing companies.

8. CONCLUSIONS AND FUTURE STEPS

After of the development of this project, I think about several positive aspects, others that can be improved and possible future steps to follow for its improvement.

As positive aspects, the following:

- It **allows reducing the subjectivity** that I mentioned in the introduction (2), since the paradigm changes: from decision-making based on intuition, to decision-making based on what the data tells us.
- The in-depth descriptive analysis carried out **has helped to better understand the employees (their opinions, sentiments and preferences) of competing companies due to the recruiting of highly qualified profiles**: what feelings their work experiences cause them; and what aspects or topics are most commented in their opinions and therefore are more relevant when assessing their work experiences. On the other hand, the comparison between companies allows us to see which sectors or group of companies are more attractive to employees and why. In addition, it allows us to delve into the weak points of each company or sector, in order to improve our employee experience and stay ahead of the competition.
- The **replicability of the project**: just as I have focused on reviews or opinions made in the most relevant work social network (Glassdoor), the techniques used can be replicated for comments written in other work networks ("Indeed", for example), surveys (the internal ones already mentioned), or internal platforms used by companies, whether their own or from external providers (Workplace, for example).

As aspects that could be improved, the following:

- I think that **preprocessing can always be improved**. In this case, find the most appropriate "lemma" or base word for each token, in order to reduce dimensionality and facilitate analysis.

- I believe that **the topic models themselves can be further fitted**, especially in cases where there are many topics of low representativeness and that share several terms (represented in pyLDAvis as small stacked bubbles). I'm talking **regarding to the models by companies**, but it is necessary more volume of reviews per company for it.

- Work on **how to carry out NLU techniques on reviews in Spanish**, with libraries that are really effective in this language.

As future steps to move forward with the project, the following:

- **Replicate the project on other networks and/or platforms**, as I mentioned.

- Solve the problem of *tediousness* involved in analyzing large volumes of text manually, through automation. Regarding this, for example, we could go further by **creating a predictive model**, that is, training a supervised classification algorithm, which adequately classifies new reviews written in real time and can predict their sentiment score with a high level of accuracy. But I consider that **it is necessary more volume of data (reviews in this case) for that.**

9. REPOSITORY

Link to the repository:

https://github.com/mcasquero/NLP_thesis.git

The repository is structured as follows:

- **Readme file** (project summary)
- **Jupyter notebooks** (which contain all the code used in the project):

(1) "Scraping and cleaning": contains the entire process of capturing data from Glassdoor and its cleaning.

(2) "Initial EDA": contains the exploratory analysis of the initial dataset previously obtained with the data collected from Glassdoor. Although the information is displayed using Python libraries, I have also created an interactive and more visual dashboard in Tableau, whose link and explanation is included in the section for this purpose (8) and in the notebook itself.

(3) "NLP": contains the rest of the project, which is everything related to Text Mining (preprocessing, 'Top Words' visualization, sentiment analysis and topic modeling).

- **Memory** (file in which the problem, questions, objectives, methodology and conclusions of the project are developed).
- **Pickle files** (contains the fitted topic models have been created)

Additionally, outside of the repository, the datasets used and collected in **.csv files** are shared through the following link to 'drive':

https://drive.google.com/drive/folders/114tHF6PMMfiT6EuLxfcsel95091kprs9?usp=sharing

In each section of the memory, it has been indicated what dataset has been used for each purpose.

10. BIBLIOGRAPHY

MCKINNEY, WES. (2013). *Python for Data Analysis.* O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

KENT D. LEE. (2014). *Python Programming Fundamentals*. Springer-Verlag London. Second Edition.

IGUAL, LAURA., SEGUI, SANTI. (2017). *Introduction to Data Science: a python approach to concepts, techniques and applications.* Springer International Publishing Switzerland.

https://www.glassdoor.es/index.htm

https://github.com/mozilla/geckodriver/releases

https://github.com/vitojph/kschool-nlp-20

https://towardsdatascience.com/

https://www.geeksforgeeks.org/

https://www.geeksforgeeks.org/nlp-gensim-tutorial-complete-guide-for-beginners/

https://realpython.com/nltk-nlp-python/

https://realpython.com/python-nltk-sentiment-analysis/

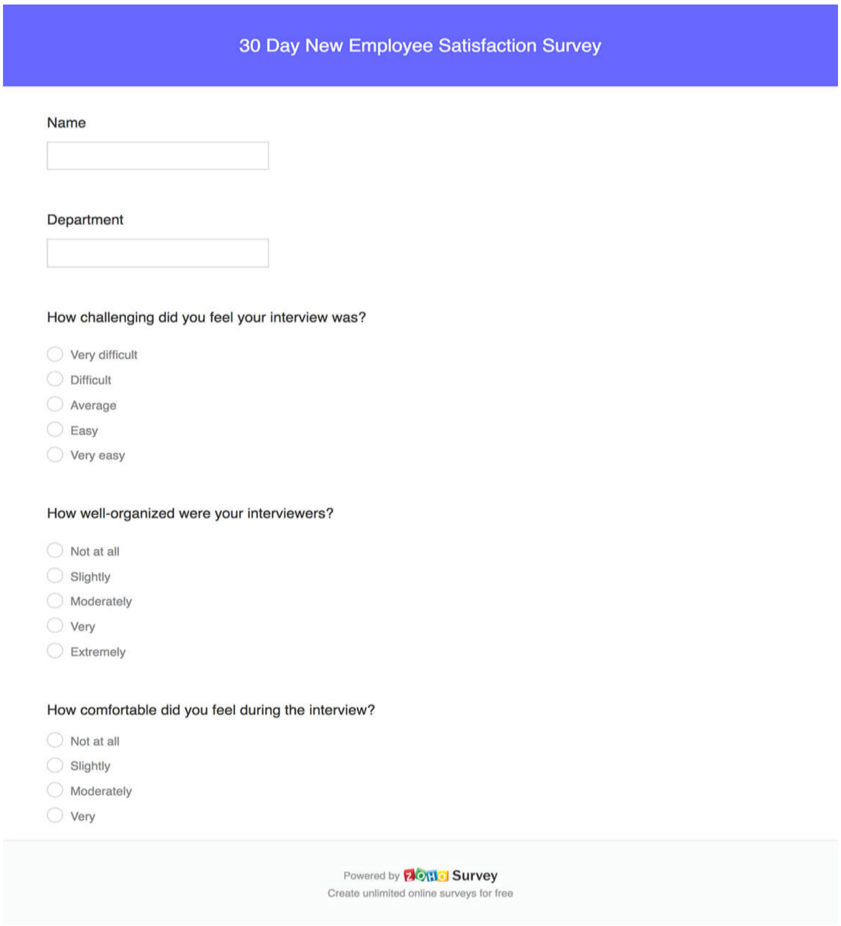
https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know

https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

11.ANNEX

[A] Examples of different surveys.

A1. Examples of different closed questions from surveys sent digitally, and after of the recruitment process:



30 Day New Employee Satisfaction Survey

Name

Department

How challenging did you feel your interview was?

○ Very difficult
○ Difficult
○ Average
○ Easy
○ Very easy

How well-organized were your interviewers?

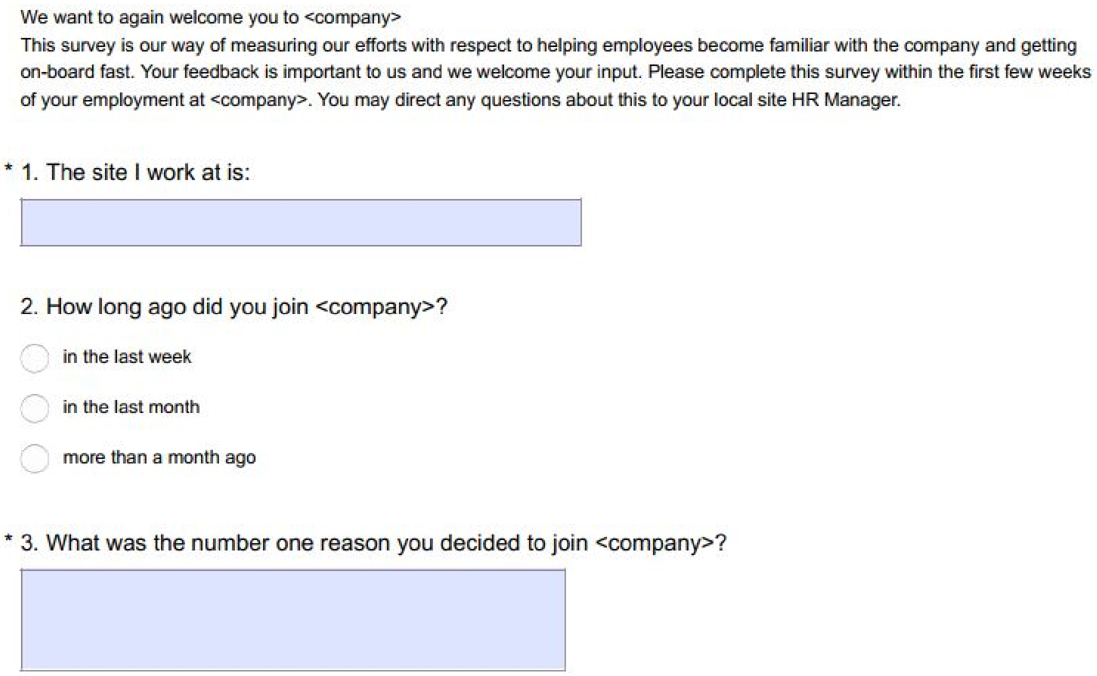
○ Not at all
○ Slightly
○ Moderately
○ Very
○ Extremely

How comfortable did you feel during the interview?

○ Not at all
○ Slightly
○ Moderately
○ Very

Powered by ZOHO Survey
Create unlimited online surveys for free

A2. Examples of open questions (question number 3) from surveys sent digitally, and after of the recruitment process:



We want to again welcome you to <company>
This survey is our way of measuring our efforts with respect to helping employees become familiar with the company and getting on-board fast. Your feedback is important to us and we welcome your input. Please complete this survey within the first few weeks of your employment at <company>. You may direct any questions about this to your local site HR Manager.

* 1. The site I work at is:

2. How long ago did you join <company>?

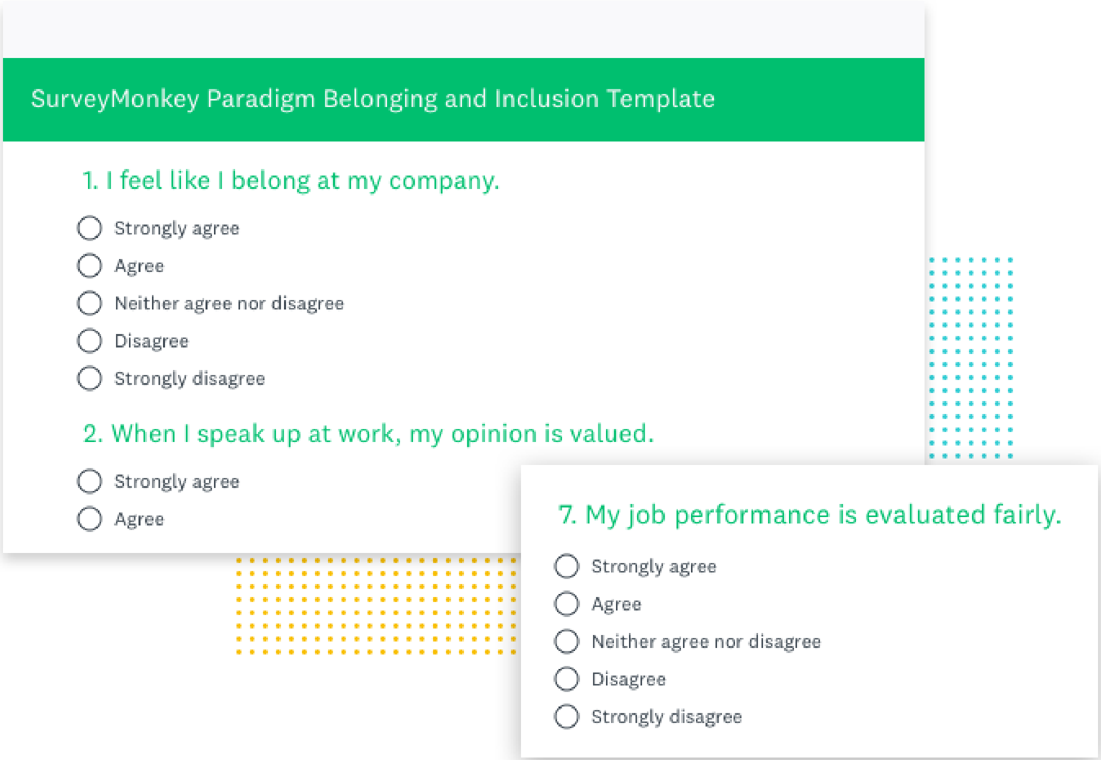
○ in the last week

○ in the last month

○ more than a month ago

* 3. What was the number one reason you decided to join <company>?

A3. Examples of different closed questions from surveys sent digitally, annually, and during the employment relationship:



SurveyMonkey Paradigm Belonging and Inclusion Template

1. I feel like I belong at my company.

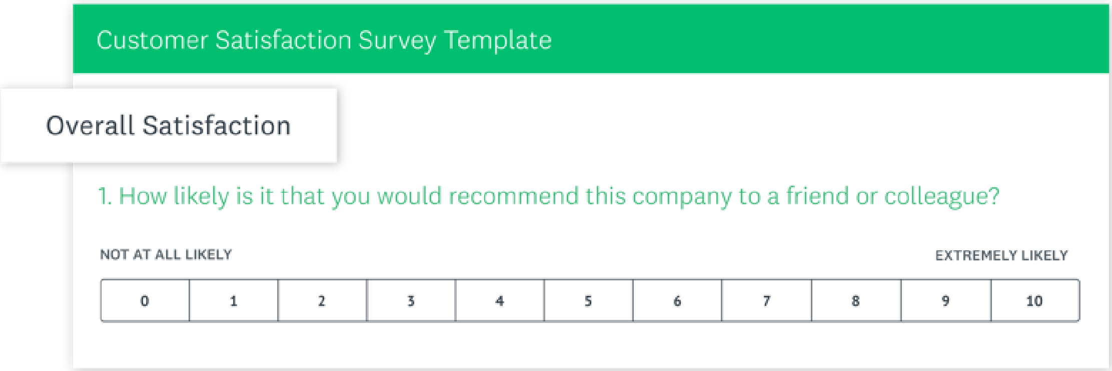
○ Strongly agree
○ Agree
○ Neither agree nor disagree
○ Disagree
○ Strongly disagree

2. When I speak up at work, my opinion is valued.

○ Strongly agree
○ Agree

7. My job performance is evaluated fairly.

○ Strongly agree
○ Agree
○ Neither agree nor disagree
○ Disagree
○ Strongly disagree



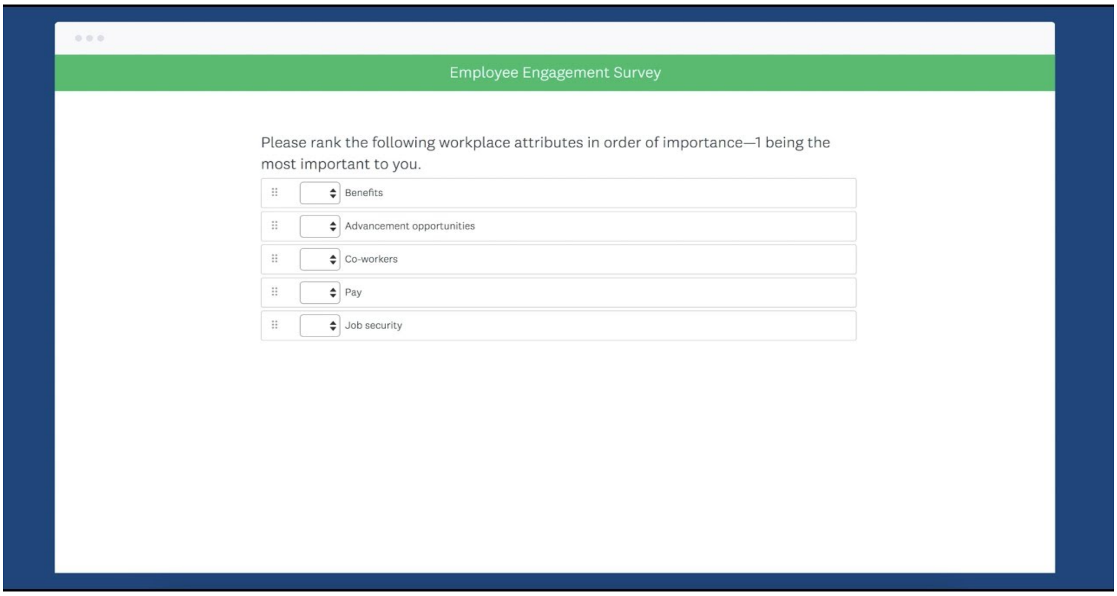
Customer Satisfaction Survey Template

Overall Satisfaction

1. How likely is it that you would recommend this company to a friend or colleague?

NOT AT ALL LIKELY EXTREMELY LIKELY

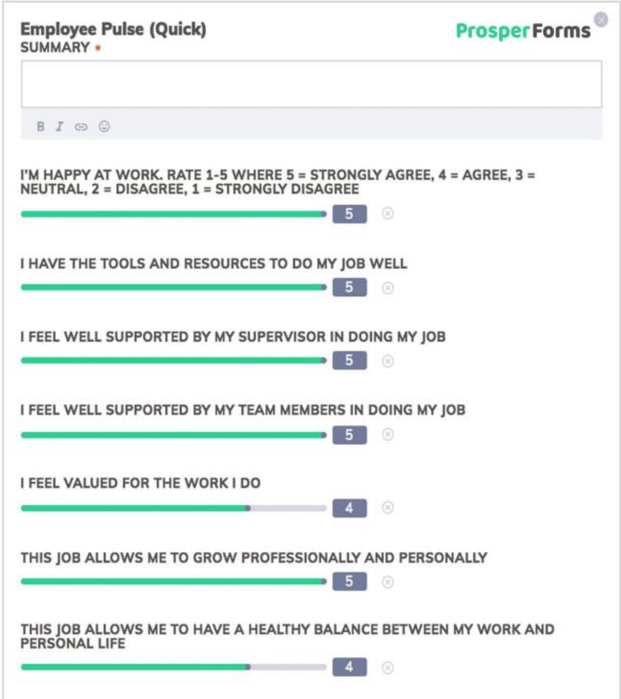
0	1	2	3	4	5	6	7	8	9	10



Employee Engagement Survey

Please rank the following workplace attributes in order of importance—1 being the most important to you.

- Benefits
- Advancement opportunities
- Co-workers
- Pay
- Job security



Employee Pulse (Quick)

SUMMARY *

I'M HAPPY AT WORK. RATE 1-5 WHERE 5 = STRONGLY AGREE, 4 = AGREE, 3 = NEUTRAL, 2 = DISAGREE, 1 = STRONGLY DISAGREE

5

I HAVE THE TOOLS AND RESOURCES TO DO MY JOB WELL

5

I FEEL WELL SUPPORTED BY MY SUPERVISOR IN DOING MY JOB

5

I FEEL WELL SUPPORTED BY MY TEAM MEMBERS IN DOING MY JOB

5

I FEEL VALUED FOR THE WORK I DO

4

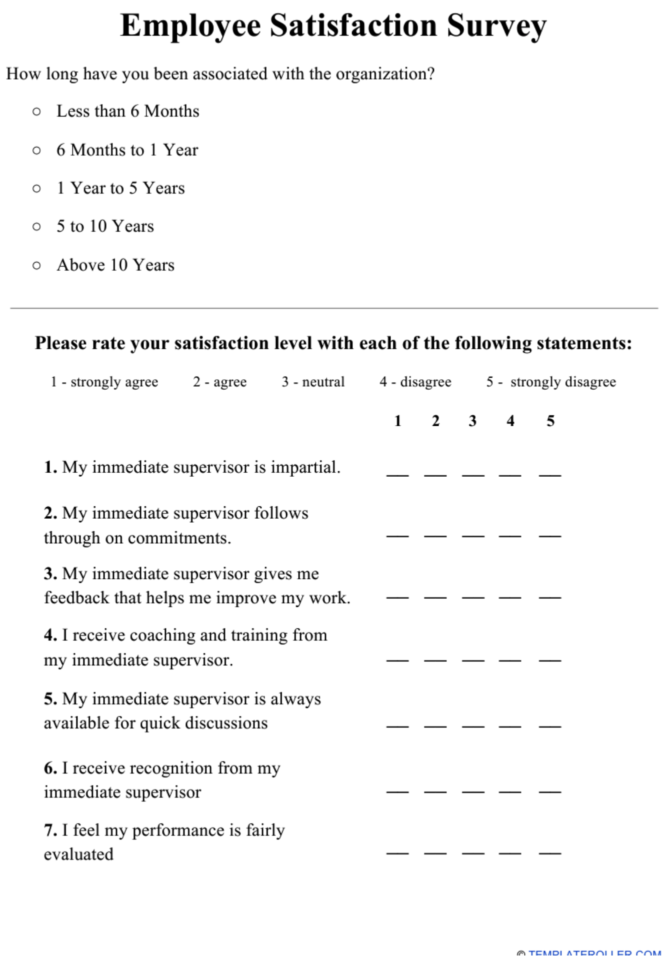
THIS JOB ALLOWS ME TO GROW PROFESSIONALLY AND PERSONALLY

5

THIS JOB ALLOWS ME TO HAVE A HEALTHY BALANCE BETWEEN MY WORK AND PERSONAL LIFE

4

A4. Examples of closed questions from surveys sent manually, annually, and during the employment relationship:



Employee Satisfaction Survey

How long have you been associated with the organization?

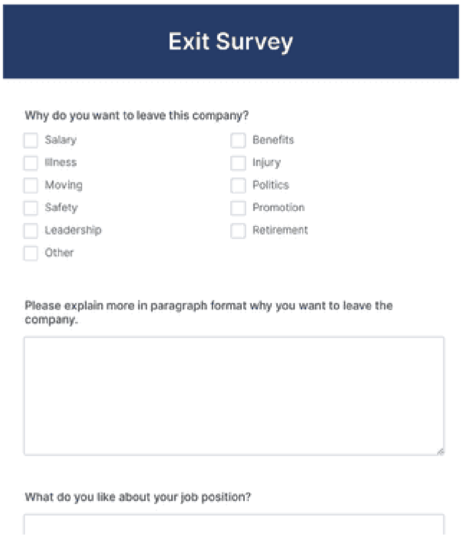
- ○ Less than 6 Months
- ○ 6 Months to 1 Year
- ○ 1 Year to 5 Years
- ○ 5 to 10 Years
- ○ Above 10 Years

Please rate your satisfaction level with each of the following statements:

1 - strongly agree 2 - agree 3 - neutral 4 - disagree 5 - strongly disagree

	1	2	3	4	5
1. My immediate supervisor is impartial.	___	___	___	___	___
2. My immediate supervisor follows through on commitments.	___	___	___	___	___
3. My immediate supervisor gives me feedback that helps me improve my work.	___	___	___	___	___
4. I receive coaching and training from my immediate supervisor.	___	___	___	___	___
5. My immediate supervisor is always available for quick discussions	___	___	___	___	___
6. I receive recognition from my immediate supervisor	___	___	___	___	___
7. I feel my performance is fairly evaluated	___	___	___	___	___

A5. Example of a mix of closed and open questions from exit simple survey sent digitally:



Exit Survey

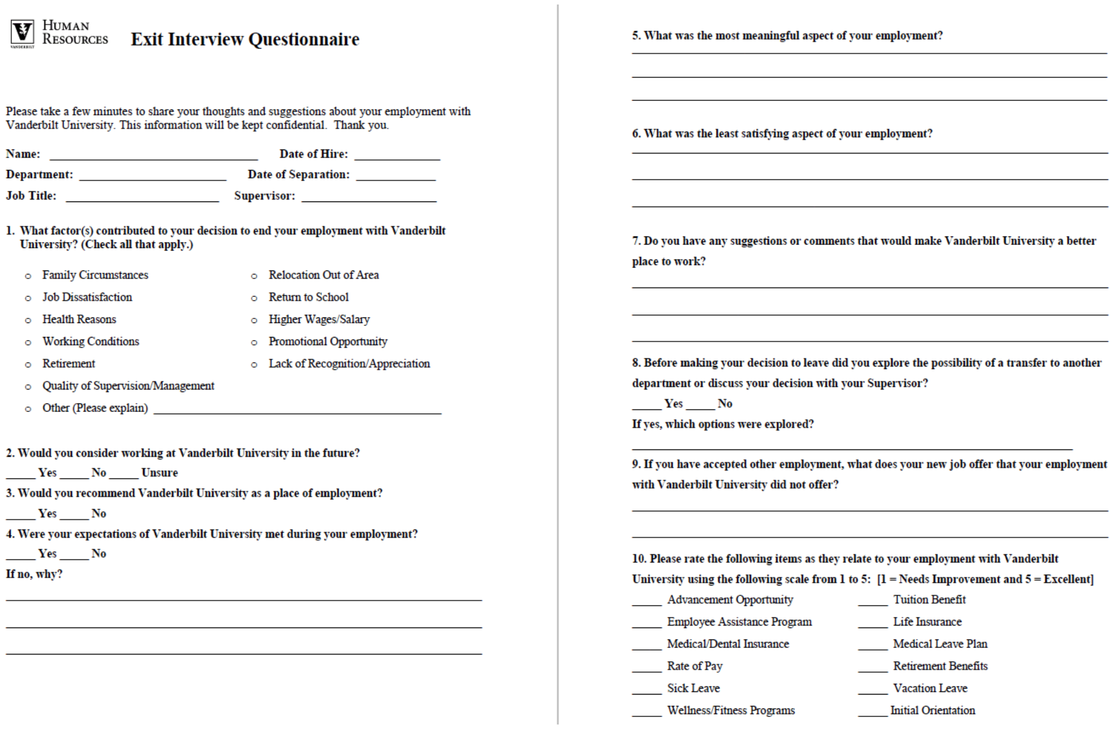
Why do you want to leave this company?

- □ Salary
- □ Illness
- □ Moving
- □ Safety
- □ Leadership
- □ Other
- □ Benefits
- □ Injury
- □ Politics
- □ Promotion
- □ Retirement

Please explain more in paragraph format why you want to leave the company.

What do you like about your job position?

A6. Example of a mix of <u>closed and open questions</u> from <u>exit survey</u> sent <u>manually</u>:



HUMAN RESOURCES

Exit Interview Questionnaire

Please take a few minutes to share your thoughts and suggestions about your employment with Vanderbilt University. This information will be kept confidential. Thank you.

Name: _____________________ **Date of Hire:** ___________

Department: ________________ **Date of Separation:** ___________

Job Title: ________________ **Supervisor:** ________________

1. What factor(s) contributed to your decision to end your employment with Vanderbilt University? (Check all that apply.)

○ Family Circumstances
○ Job Dissatisfaction
○ Health Reasons
○ Working Conditions
○ Retirement
○ Quality of Supervision/Management
○ Other (Please explain) ______________________________
○ Relocation Out of Area
○ Return to School
○ Higher Wages/Salary
○ Promotional Opportunity
○ Lack of Recognition/Appreciation

2. Would you consider working at Vanderbilt University in the future?
_____ **Yes** _____ **No** _____ **Unsure**

3. Would you recommend Vanderbilt University as a place of employment?
_____ **Yes** _____ **No**

4. Were your expectations of Vanderbilt University met during your employment?
_____ **Yes** _____ **No**

If no, why?

5. What was the most meaningful aspect of your employment?

6. What was the least satisfying aspect of your employment?

7. Do you have any suggestions or comments that would make Vanderbilt University a better place to work?

8. Before making your decision to leave did you explore the possibility of a transfer to another department or discuss your decision with your Supervisor?
_____ **Yes** _____ **No**

If yes, which options were explored?

9. If you have accepted other employment, what does your new job offer that your employment with Vanderbilt University did not offer?

10. Please rate the following items as they relate to your employment with Vanderbilt University using the following scale from 1 to 5: [1 = Needs Improvement and 5 = Excellent]

_____ Advancement Opportunity
_____ Employee Assistance Program
_____ Medical/Dental Insurance
_____ Rate of Pay
_____ Sick Leave
_____ Wellness/Fitness Programs
_____ Tuition Benefit
_____ Life Insurance
_____ Medical Leave Plan
_____ Retirement Benefits
_____ Vacation Leave
_____ Initial Orientation