UNIVERSITÀ DI PISA

BUSINESS & PROJECT MANAGEMENT

< Sentiment Analysis >

Project made by**:**

- **Domenico Armillotta -** badge : 643020 - email: d.armillotta@studenti.unipi.it
- **Leonardo Bellizzi -** badge : 643019 email: l.bellizzi@studenti.unipi.it
- **Valerio Secondulfo -** badge : 643022 email: val.secondulfo@studenti.unipi.it

PROJECT GOAL AND RESEARCH QUESTIONS

The research will focus on the topic of **consumer sentiment**, focusing on **sentiment analysis**, i.e. identification, extract, quantify, and study affective states and subjective information using natural language processing and text analysis.

First Question:

The research will focus on the following question:

"Can sentiment analysis technology be used in a realistic business environment to understand customer feedback left on social media?"

Assuming that the following program is within a CRM.

Then we will analyze some textual reviews and we will label them with our rating model, and we will understand how much the rating of the classified reviews differs from the automatic system, from the rating

of real reviews (assuming that the stars left by users are balanced

with the textual review left).

Second Question:

The second research focuses on the following question:

“Which are the main topics of the sentiment analysis application in the management field?”

The use of sentiment analysis is imperative nowadays for a business company in order to retrieve a customer’s feedback in an easy and rapid way but, along with that, a company must always keep in touch and grow along with the evolution of technology. We focused our attention in the detection of which are the main uses and possibilities of the application of the sentiment analysis through the analysis,conducted by using Natural Language Processing techniques, of papers taken from MIT Sloan school related to the management and business field.

PROCESS DESCRIPTION

First Question:

The process of answering the first question is not very short, but we wanted to get realistic results.

Our goal is to understand how much the results extracted from the automatic sentiment analysis differ from the real feelings left in the online reviews, in this case hotel reviews.

After training several classifiers and choosing the best one, the comments will be automatically classified with the npl technique in Good, Neutral, Bad and then they will be compared with the real ratings of the reviews.

All the technical details and the various steps will be shown in the “Technical Report” below.

Second Question:

The process of the second question consists in the use of the natural processing language on a series of papers in the field of management taken from MIT’s management school.

After a first conversion of the papers,scraped in a .rcs format, into a .csv we searched among them the ones related to our field of interest by selecting only those containing information about data mining and sentiment analysis field through the use of the Python language.

Once we’ve obtained the papers of our interest we then applied a text pre-processing phase and finally used NLP techniques in order to obtain readable diagrams containing the main topics inside the documents.

In particular we first analyzed each paper alone in order to always have an idea of the relation of sentiment analysis and the business field through time and at the end we considered all the papers we analyzed as one and repeated the analysis in order to retrieve the general topics.

TECHNICAL REPORT

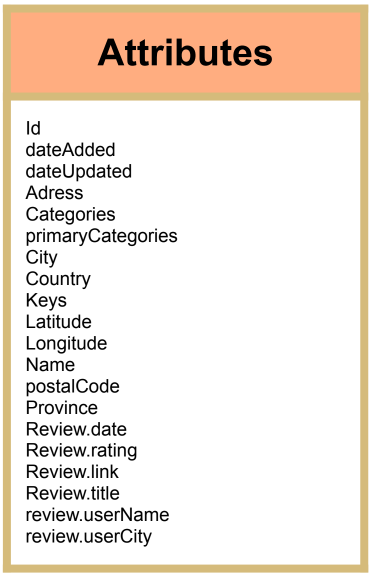
First Question

1. Dataset:

Our dataset was taken by Kaggle. The dataset collects all the reviews of various hotels from various online sites , therefore from different sources.

Link:https://www.kaggle.com/datafiniti/hotel-reviews?select=Datafiniti_Hotel_Reviews.csv

Attribute : 25
Dimension : 30 000 reviews



Attributes

Id
dateAdded
dateUpdated
Adress
Categories
primaryCategories
City
Country
Keys
Latitude
Longitude
Name
postalCode
Province
Review.date
Review.rating
Review.link
Review.title
review.userName
review.userCity

2. Dataset Splitting & Label Phase:

In order to evaluate the results of the different classifiers we need to split the dataset into training set and test set, in this case 75% train and 25% test.

This phase is important for the training phase of the classifiers. In this case we have made an important assumption, which is that most of the reviews coincide with the stars given in the review. We have labeled our comments following this pattern:

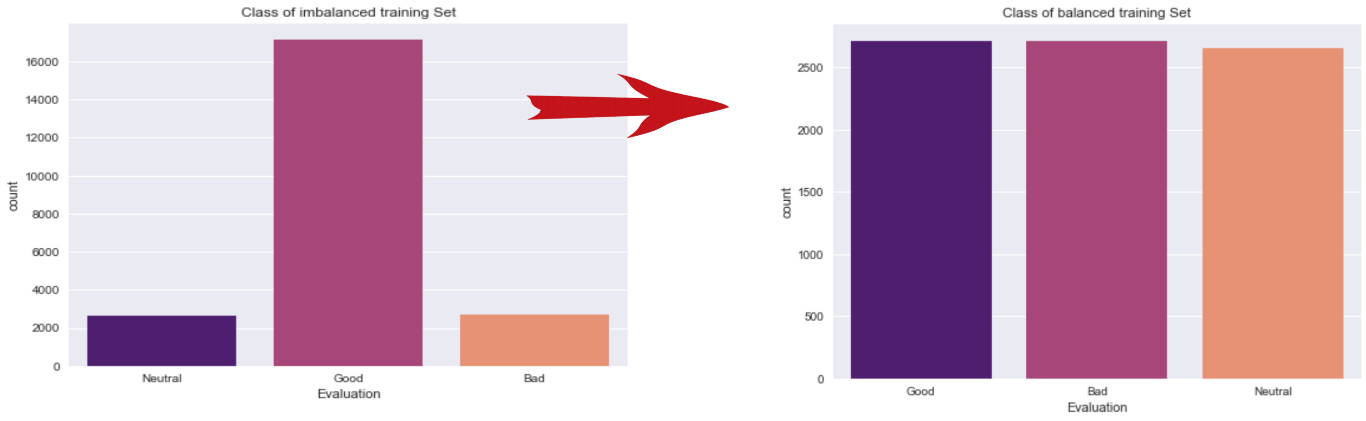
- 4-5 stars⟶ Good
- 3 stars ⟶ Neutral
- 1-2 stars ⟶ Bad

Undersampling Phase:

After the label phase we were able to see how was the distribution of the data for the training set. We noticed that our training set was very unbalanced , and that's not good for classifier training at all, so we decided to undersampling on the training set, and thus obtaining:

8072 comment labeled:

- 2711 Rating Good
- 2711 Rating Neutral
- 2653 Rating Bad

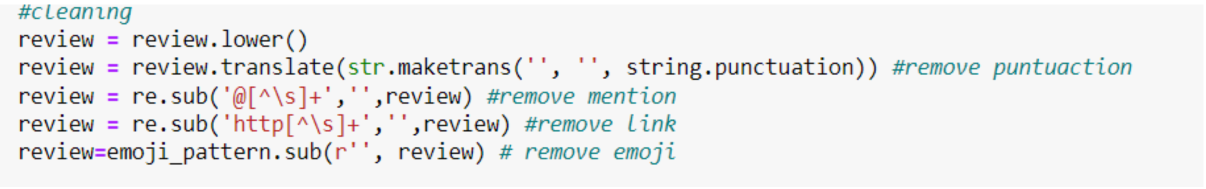


-

3. Cleaning Phase

In order to obtain a cleaned text to pass to Preprocessing phase we did steps like:

1. Lower case words
2. Remove punctuation from sentences
3. Remove mention @ or Hashtags from sentences
4. Remove link http from sentences
5. Remove emoji from sentences



```
#cleaning
review = review.lower()
review = review.translate(str.maketrans('', '', string.punctuation)) #remove puntuaction
review = re.sub('@[^\s]+','',review) #remove mention
review = re.sub('http[^\s]+','',review) #remove link
review=emoji_pattern.sub(r'', review) # remove emoji
```

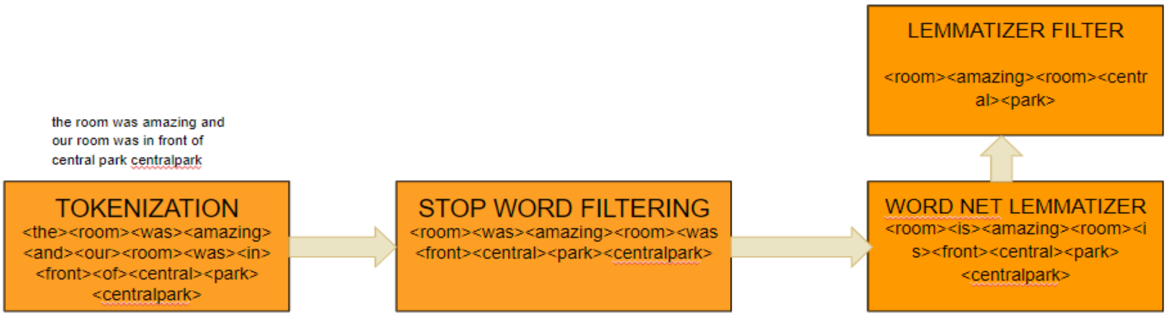
4. Text Elaboration

We performed pre-processing steps to transform our comments into TF IDF representation in Python. To carry out them, we used the CountVectorizer function from sklearn.

- **Tokenization**: in this first step we transformed a stream of characters into a stream of processing units called tokens. In this way each text is represented as a set of words. To do this we set some parameters of the CountVectorizer function
- **Stop Words Filtering**: in this second step we removed the stop words (which provide little or no useful information to the text analysis) setting the stop_words parameter equal to 'english
- **Lemmatization**: is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form.Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence. To do this task we use the library nltk.

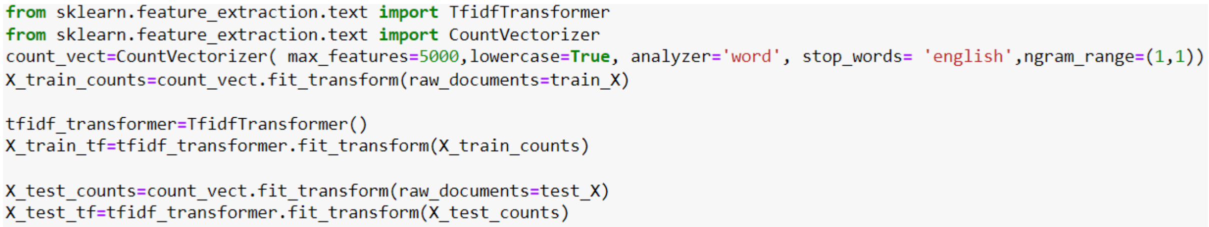
-> better : good

- **Lemmatization Filtering**: in this fourth step we reduced the number of stems maintaining only the most relevant ones. We set max_features=5000 as CountVectorizer's parameter which considers the top max_features ordered by term frequency across the corpus.


the room was amazing and
our room was in front of
central park centralpark
TOKENIZATION
<the><room><was><amazing><and><our><room><was><in><front><of><central><park><centralpark>
STOP WORD FILTERING
<room><was><amazing><room><was<front><central><park><centralpark>
WORD NET LEMMATIZER
<room><is><amazing><room><is><front><central><park><centralpark>
LEMMATIZER FILTER
<room><amazing><room><central><park>

5. Feature extraction (TF-IDF & N-Grams)

TF-IDF was used as a feature extraction technique. Was used to measure the importance of a term with respect to a document or collection of documents. This function increases proportionally to the number of times the term is contained in the document, but increases inversely with the frequency of the term in the collection. The idea behind this behavior is to give more importance to the terms that appear in the document, but which in general are infrequent



```
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import CountVectorizer
count_vect=CountVectorizer( max_features=5000,lowercase=True, analyzer='word', stop_words= 'english',ngram_range=(1,1))
X_train_counts=count_vect.fit_transform(raw_documents=train_X)

tfidf_transformer=TfidfTransformer()
X_train_tf=tfidf_transformer.fit_transform(X_train_counts)

X_test_counts=count_vect.fit_transform(raw_documents=test_X)
X_test_tf=tfidf_transformer.fit_transform(X_test_counts)
```

With TF-IDF the N-GRAM was used. N-grams are continuous sequences of words or symbols or tokens in a document. In technical terms, they can be defined as the neighboring sequences of items in a document.

6. Classifiers used

Classifiers that have been used are:

- Multinomial Naive Bayes;
- Support Vector Machine;
- Random Forest;
- Decision Tree;
- KNN;
- Bagging;
- AdaBoost.

7. Cross validation & choose the best Classifier

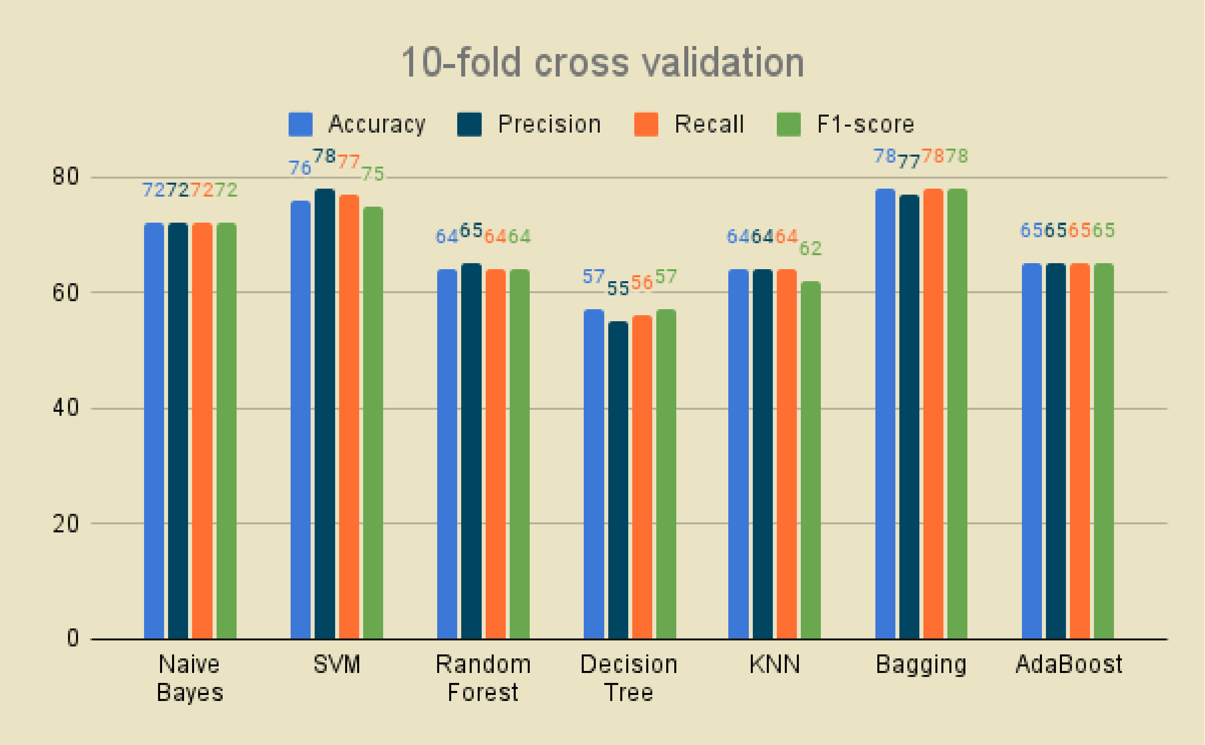
Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. In particular it was used to figure out the performance of the classifiers and understand if classifiers were overfitting and detect the best one.

We used the K-FOLD and not the STRATIFIED K-FOLD because our data are balanced.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. In our case the chosen K is equal to 10

Cross-validation was primarily used to estimate the skill of a machine learning model on unseen data.

Here the results of our classifiers under the 10-Fold Cross Validation.

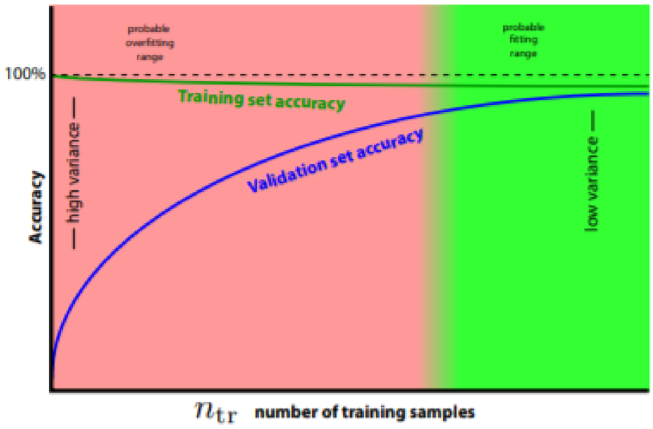


Attention to Overfitting classifiers

We tried to modify the parameters to get better results, but plotting the learning curve, we got suspicious seeing the training accuracy so far from the validation accuracy, and doing some research we realized that two of our classifiers were overfitting. So we changed the parameters to try to avoid overfitting.

The tuning work for RANDOM FOREST even if we lost some accuracy score but DECISION TREE was still in an overfitting situation.

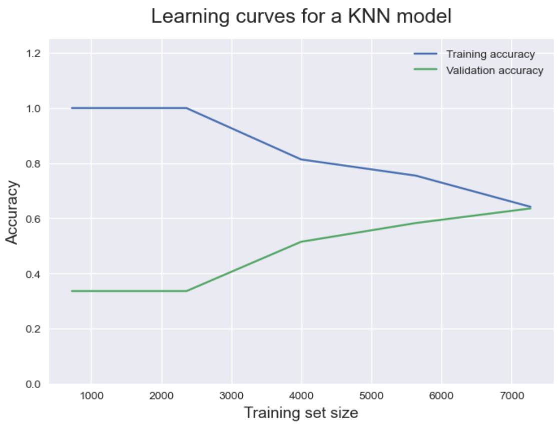
According to **Tom Mitchell, Machine Learning, McGraw-Hill Science/Engineering/Math; 1997**, the Overfitting situation is caught in the following scheme:

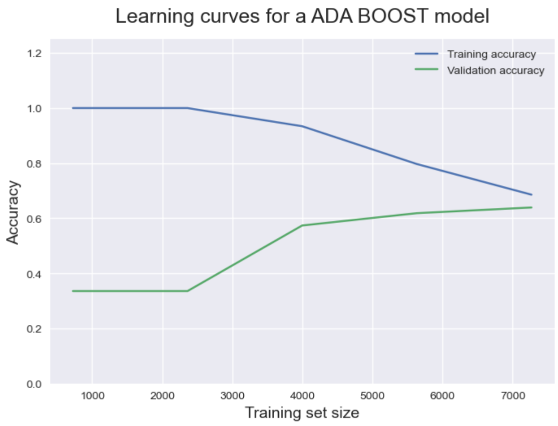


So the best classifiers that are not overfitted are **Bagging** and **SVM** and are the classifiers that have best balance scores in terms of Accuracy, Precision, Recall and F1 .

Basing only on accuracy score isn't always a good practice. Accuracy give us only a general idea on how the classifiers worked.

Here the learning curves of the two best classifiers:

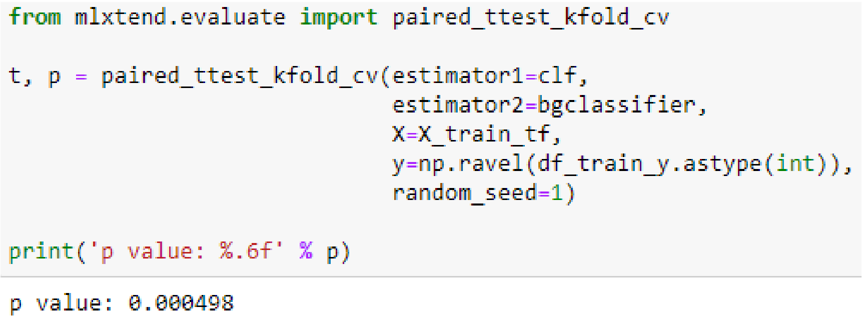




Picking the best model

We used the t test to compare the two best classifiers obtained from the k fold, because we may have gotten the same model from two different classifiers.

And these are the results:



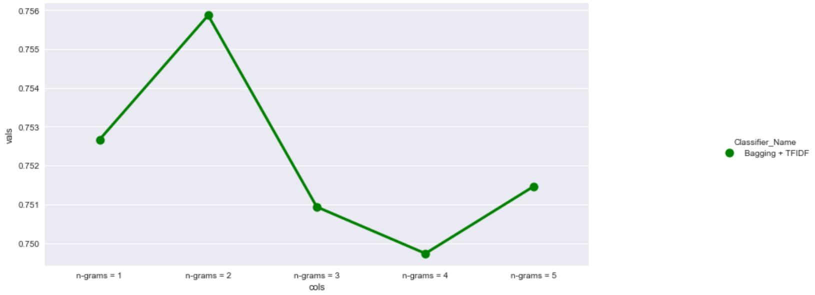
```
from mlxtend.evaluate import paired_ttest_kfold_cv

t, p = paired_ttest_kfold_cv(estimator1=clf,
                             estimator2=bgclassifier,
                             X=X_train_tf,
                             y=np.ravel(df_train_y.astype(int)),
                             random_seed=1)

print('p value: %.6f' % p)
```

```
p value: 0.000498
```

Since p<α (=0.05), we can accept the Alternative Hypothesis. This means that we take the best classifier between the two to build our application. **Bagging Classifier**

8. Bagging Classifier Insight

We tried to use different N-Grams to realize if our classifier accuracy could have been improved.



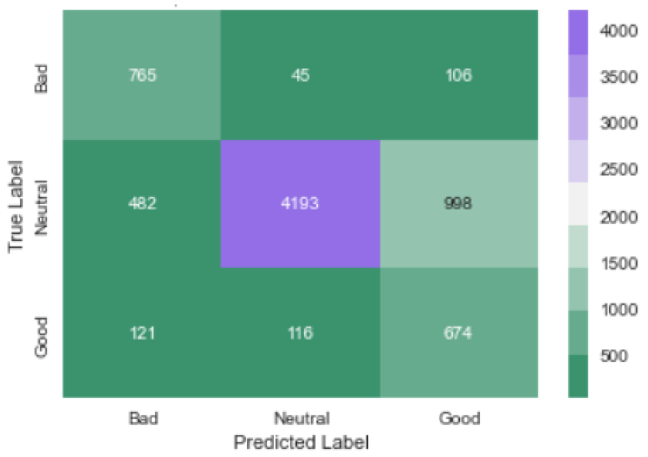
With N-Gram = (1,2) the performance of **Bagging** increases.

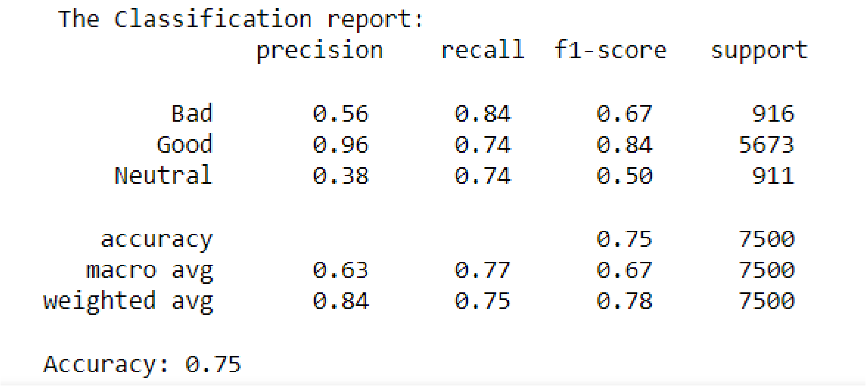
9. Final Classifier

Here is the confusion matrix of the Bagging Classifier.

This grader will be used in the app for rating reviews as good, neutral and bad.

The accuracy is not very high, but it was the maximum obtainable in this situation, because otherwise we would have risked the overfitting of the model, as seen in the previous graphs.

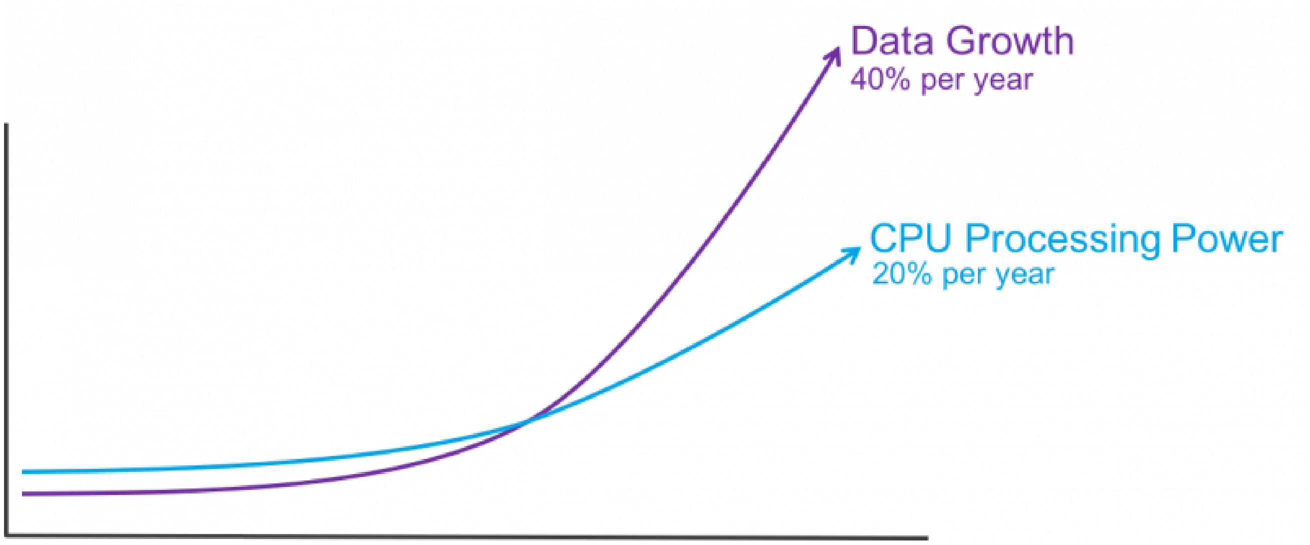




```
 The Classification report:
              precision    recall  f1-score   support

         Bad       0.56      0.84      0.67       916
        Good       0.96      0.74      0.84      5673
     Neutral       0.38      0.74      0.50       911

    accuracy                           0.75      7500
   macro avg       0.63      0.77      0.67      7500
weighted avg       0.84      0.75      0.78      7500

Accuracy: 0.75
```

10. NEURAL NETWORK CLASSIFICATION

Neural networks usually require much more data than traditional machine learning algorithms, as in at least thousands if not millions of labeled samples. This isn't an easy problem to deal with and many machine learning problems can be solved well with less data if you use other algorithms. Although there are some cases where neural networks do well with little data, most of the time they don't. In this case, a simple algorithm like naive Bayes, which deals much better with little data, would be the appropriate choice.



Usually, neural networks are also more computationally expensive than traditional algorithms. State of the art deep learning algorithms, which realize successful training of really deep neural networks, can take several weeks to train completely from scratch. By contrast, most traditional machine learning algorithms take much less time to train, ranging from a few minutes to a few hours or days.

The amount of computational power needed for a neural network depends heavily on the size of your data, but also on the depth and complexity of your network. For example, a neural network with one layer and 50 neurons will be much faster than a random forest with 1,000 trees. By comparison, a neural network with 50 layers will be much slower than a random forest with only 10 trees.

Given the aforementioned reasons we prefered to not use this approach and use traditional ML algorithms that work well with small data

Second Question:

1. Dataset:

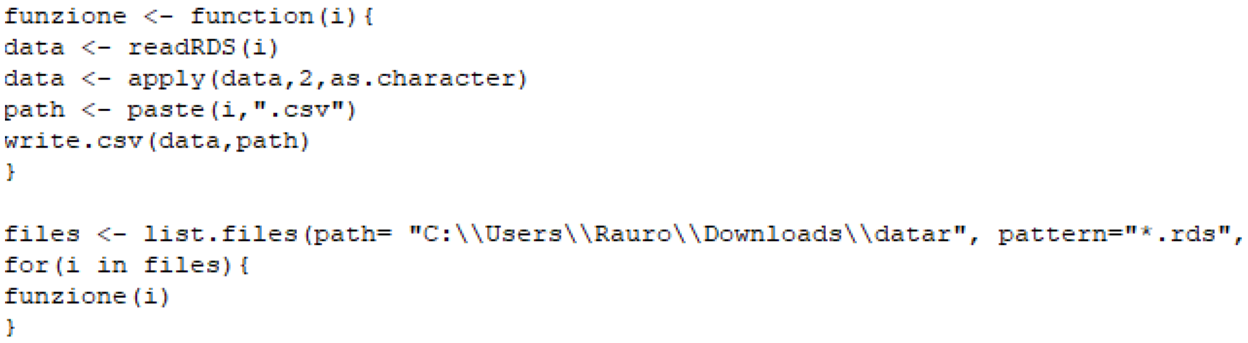
The Dataset used for this section consists in 128 papers taken from the MIT Sloan management school covering the years from 1990 to 2022.
For each year there are four papers, one for each season (spring,summer,fall and winter).

Inside each paper there are multiple articles related to many different topics about business and economy divided as:

Link	Full page	Title	Author	Year	Text	References	About

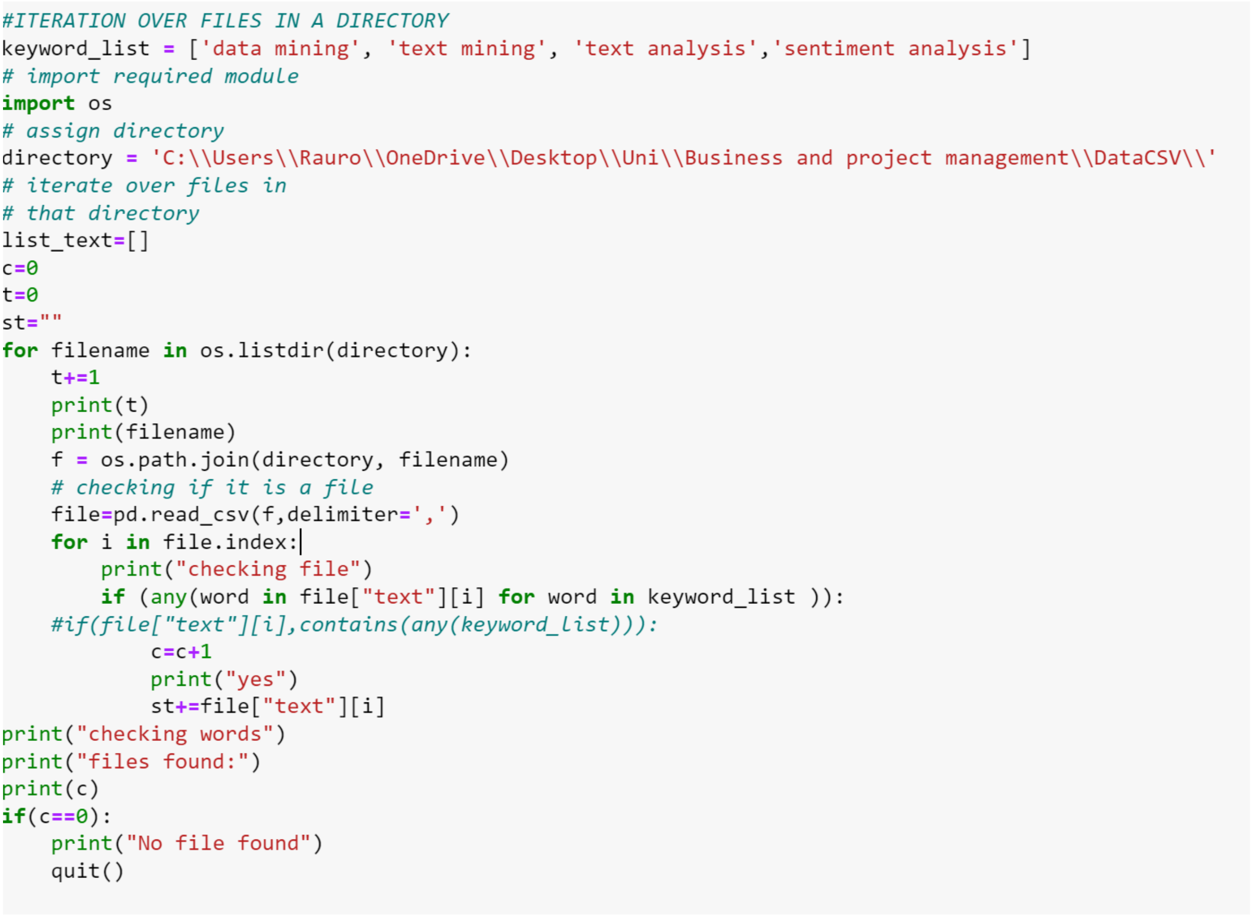
2. Conversion and filtering:

The files scraped were in the RDS format so before any other operation we had to convert them into CSV format.
In order to do that we used a script in R which took the whole folder containing the papers and converted each of them into the requested format as shown below.



```
funzione <- function(i){
data <- readRDS(i)
data <- apply(data,2,as.character)
path <- paste(i,".csv")
write.csv(data,path)
}

files <- list.files(path= "C:\\Users\\Rauro\\Downloads\\datar", pattern="*.rds",
for(i in files){
funzione(i)
}
```

After the conversion we had to analyze the files in order to see if there were present articles related to our topic research.
In order to do so, we used a python script (as shown below) that scanned each paper and searched inside of them the presence of articles related to the Data Mining and Sentiment Analysis field.



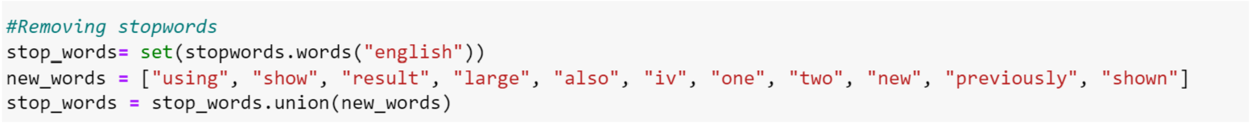
```
#ITERATION OVER FILES IN A DIRECTORY
keyword_list = ['data mining', 'text mining', 'text analysis','sentiment analysis']
# import required module
import os
# assign directory
directory = 'C:\\Users\\Rauro\\OneDrive\\Desktop\\Uni\\Business and project management\\DataCSV\\'
# iterate over files in
# that directory
list_text=[]
c=0
t=0
st=""
for filename in os.listdir(directory):
    t+=1
    print(t)
    print(filename)
    f = os.path.join(directory, filename)
    # checking if it is a file
    file=pd.read_csv(f,delimiter=',')
    for i in file.index:
        print("checking file")
        if (any(word in file["text"][i] for word in keyword_list )):
    #if(file["text"][i],contains(any(keyword_list))):
            c=c+1
            print("yes")
            st+=file["text"][i]
print("checking words")
print("files found:")
print(c)
if(c==0):
    print("No file found")
    quit()
```

If a file is related to our field of interest the script will append it inside a string.
At the end of the process we obtained a string containing all the articles related to the sentiment analysis.
Of over 128 papers we found about 25 articles.

In case the articles found would have been more than a certain threshold we would have to create a new .txt file containing all the documents to work on.

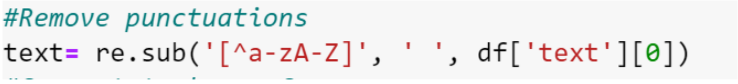
3. Text pre-processing:

On the string obtained we then proceeded to apply a text-processing phase.
Before applying the text processing we created a small set of new stop words to add to the one provided by the nltk python library.



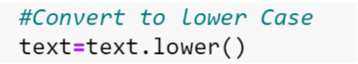
```
#Removing stopwords
stop_words= set(stopwords.words("english"))
new_words = ["using", "show", "result", "large", "also", "iv", "one", "two", "new", "previously", "shown"]
stop_words = stop_words.union(new_words)
```

Then we removed the punctuation



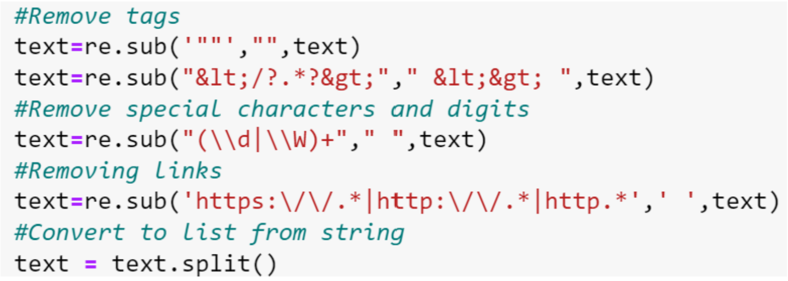
```
#Remove punctuations
text= re.sub('[^a-zA-Z]', ' ', df['text'][0])
```

Converted the whole text to lowercase



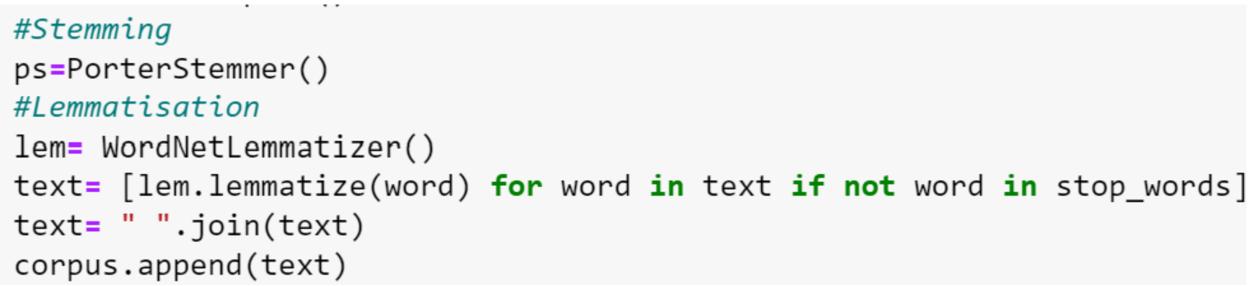
```
#Convert to Lower Case
text=text.lower()
```

We then removed tags, special characters, digits,links and converted the whole string to a list.



```
#Remove tags
text=re.sub('""',"",text)
text=re.sub("&lt;/?.*?&gt;"," &lt;&gt; ",text)
#Remove special characters and digits
text=re.sub("(\\d|\\W)+"," ",text)
#Removing links
text=re.sub('https:\/\/.*|http:\/\/.*|http.*',' ',text)
#Convert to list from string
text = text.split()
```

After those processes we then applied Stemming and Lemmatization to the text



```
#Stemming
ps=PorterStemmer()
#Lemmatisation
lem= WordNetLemmatizer()
text= [lem.lemmatize(word) for word in text if not word in stop_words]
text= " ".join(text)
corpus.append(text)
```

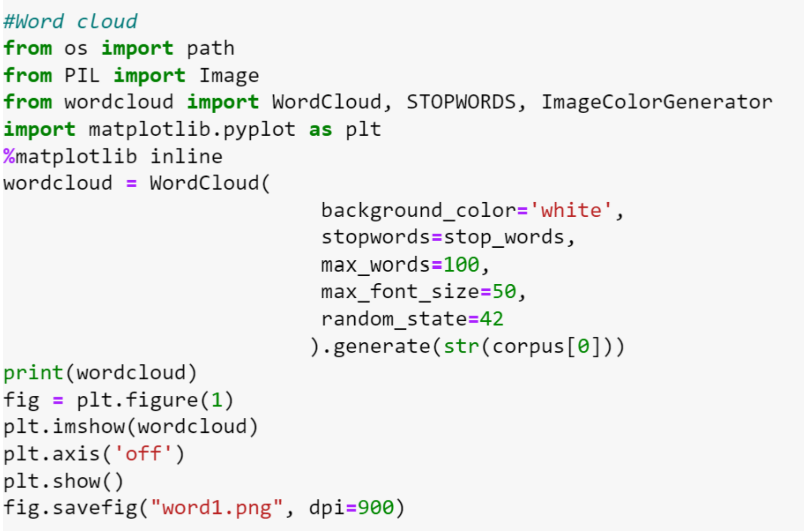
Stemming is applied to remove suffixes from the words, whereas Lemmatization is used to group together words by their inflicted form.

4. Topic extraction:

Once that the text pre-processing has been applied we can than proceed to the topic extraction.

For this part we considered four different approaches based on the use of natural language processing.

The first approach used is the Word Cloud with which we can have an image showing the most frequently used words inside the document based on the size of the text itself. The bigger the text the higher is the frequency of a single word.
(The result obtained will be shown in the result section).



```
#Word cloud
from os import path
from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import matplotlib.pyplot as plt
%matplotlib inline
wordcloud = WordCloud(
                          background_color='white',
                          stopwords=stop_words,
                          max_words=100,
                          max_font_size=50,
                          random_state=42
                         ).generate(str(corpus[0]))
print(wordcloud)
fig = plt.figure(1)
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
fig.savefig("word1.png", dpi=900)
```

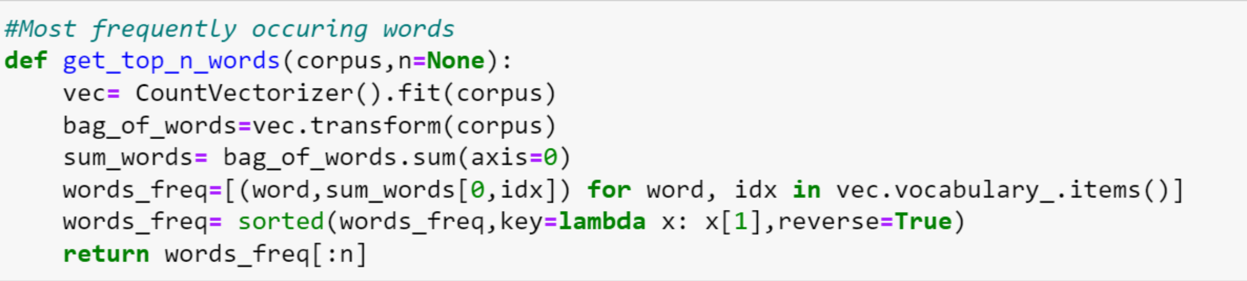
After this we proceeded to create three different n-grams showing the frequency of n group of words inside the document.

In order to do so we prepared the text in order to be recognized by the machine learning algorithms by tokenizing and vectorizing it.

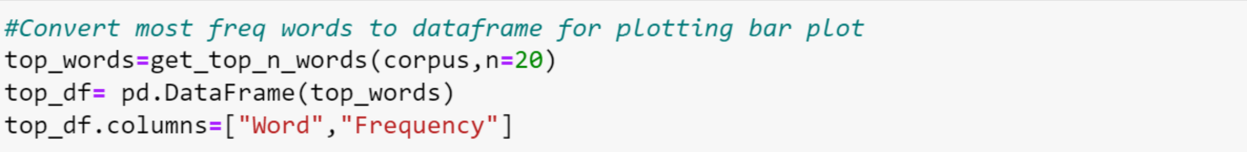
The tokenization transforms the continuous text into a list of single words, the vectorization will then convert the list into a matrix of integers.

For doing so we used the bag of word model which ignores the sequence of words and only considers their frequency.

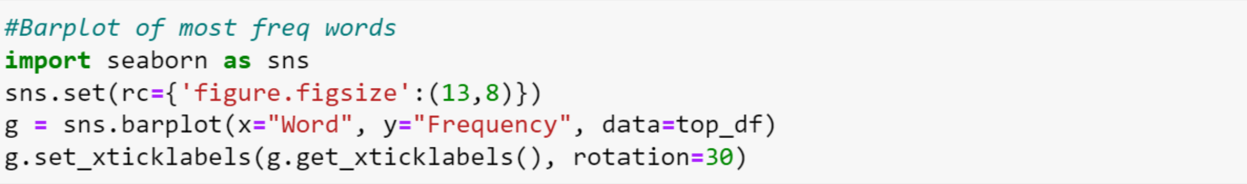
One word n-gram.



```
#Most frequently occuring words
def get_top_n_words(corpus,n=None):
    vec= CountVectorizer().fit(corpus)
    bag_of_words=vec.transform(corpus)
    sum_words= bag_of_words.sum(axis=0)
    words_freq=[(word,sum_words[0,idx]) for word, idx in vec.vocabulary_.items()]
    words_freq= sorted(words_freq,key=lambda x: x[1],reverse=True)
    return words_freq[:n]
```

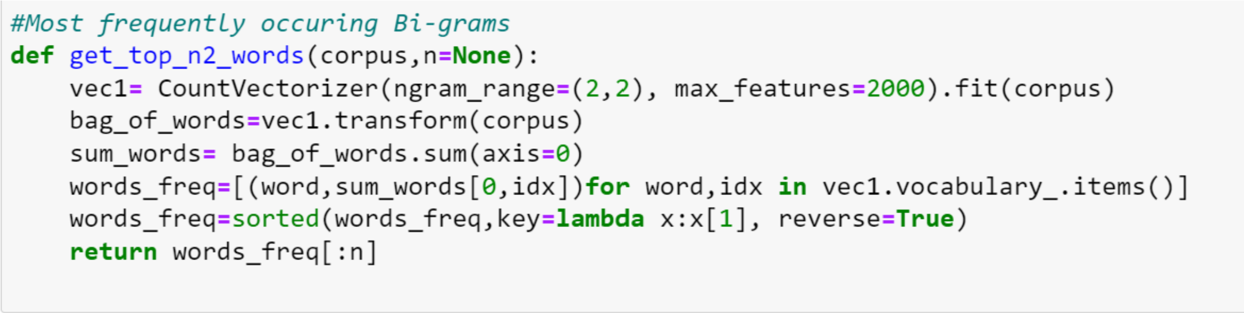


```
#Convert most freq words to dataframe for plotting bar plot
top_words=get_top_n_words(corpus,n=20)
top_df= pd.DataFrame(top_words)
top_df.columns=["Word","Frequency"]
```

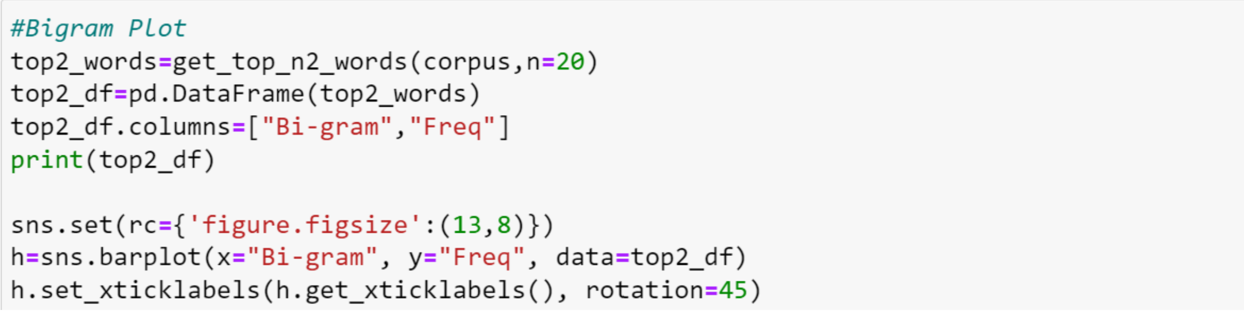


```
#Barplot of most freq words
import seaborn as sns
sns.set(rc={'figure.figsize':(13,8)})
g = sns.barplot(x="Word", y="Frequency", data=top_df)
g.set_xticklabels(g.get_xticklabels(), rotation=30)
```

Two words n-gram

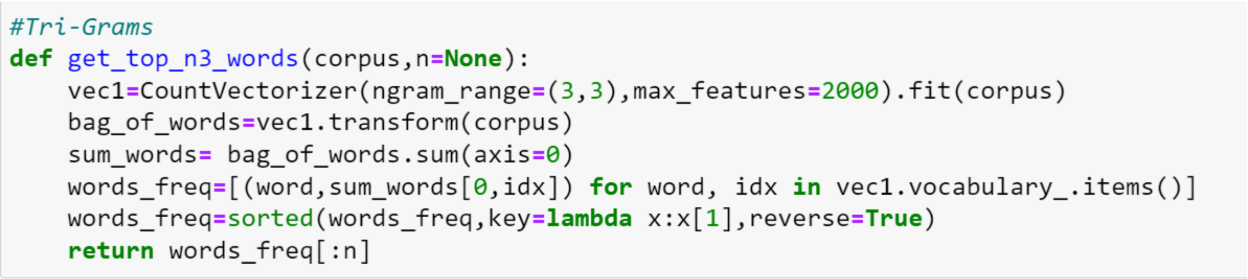


```
#Most frequently occuring Bi-grams
def get_top_n2_words(corpus,n=None):
    vec1= CountVectorizer(ngram_range=(2,2), max_features=2000).fit(corpus)
    bag_of_words=vec1.transform(corpus)
    sum_words= bag_of_words.sum(axis=0)
    words_freq=[(word,sum_words[0,idx])for word,idx in vec1.vocabulary_.items()]
    words_freq=sorted(words_freq,key=lambda x:x[1], reverse=True)
    return words_freq[:n]
```

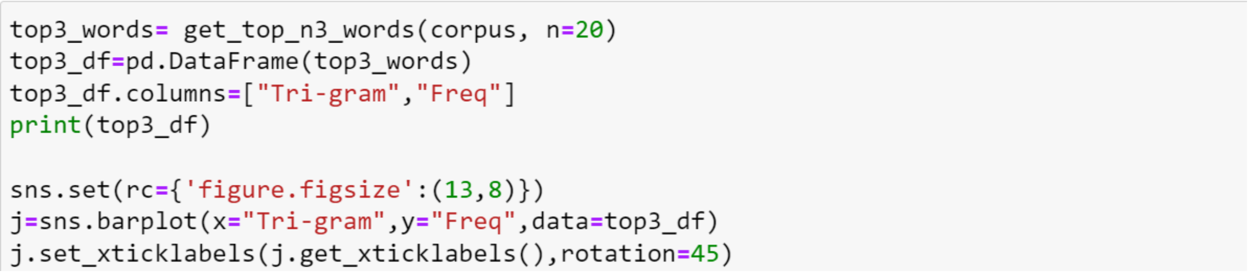


```
#Bigram Plot
top2_words=get_top_n2_words(corpus,n=20)
top2_df=pd.DataFrame(top2_words)
top2_df.columns=["Bi-gram","Freq"]
print(top2_df)

sns.set(rc={'figure.figsize':(13,8)})
h=sns.barplot(x="Bi-gram", y="Freq", data=top2_df)
h.set_xticklabels(h.get_xticklabels(), rotation=45)
```

Three words n-gram



```python
#Tri-Grams
def get_top_n3_words(corpus,n=None):
    vec1=CountVectorizer(ngram_range=(3,3),max_features=2000).fit(corpus)
    bag_of_words=vec1.transform(corpus)
    sum_words= bag_of_words.sum(axis=0)
    words_freq=[(word,sum_words[0,idx]) for word, idx in vec1.vocabulary_.items()]
    words_freq=sorted(words_freq,key=lambda x:x[1],reverse=True)
    return words_freq[:n]
```



```python
top3_words= get_top_n3_words(corpus, n=20)
top3_df=pd.DataFrame(top3_words)
top3_df.columns=["Tri-gram","Freq"]
print(top3_df)

sns.set(rc={'figure.figsize':(13,8)})
j=sns.barplot(x="Tri-gram",y="Freq",data=top3_df)
j.set_xticklabels(j.get_xticklabels(),rotation=45)
```

The last npl process applied to the text is the TF-IDF vectorization.
CountVectorizer focuses on the frequency of the words. By doing so it could happen that words with more important meaning but lower frequency may be discarded by the algorithm. By applying the TF-IDF vectorization we penalize in this case the words that appear several times in the document highlighting the ones which are more important to the context.

TF determines the Term Frequency of a single term.

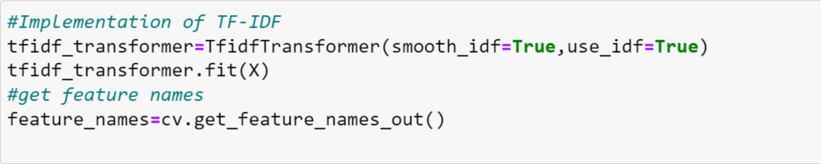
IDF determines the inverse document frequency.

TF= $Frequency\ of\ a\ single\ term/total\ number\ of\ terms\ in\ the\ document$

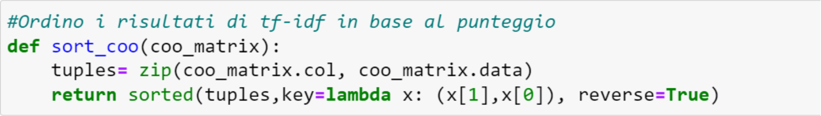
$$TF = \frac{Frequency\ of\ a\ single\ term}{Total\ number\ of\ terms\ in\ the\ document}$$

$$IDF = \frac{log(Total\ documents)}{Number\ of\ documents\ with\ the\ term}$$

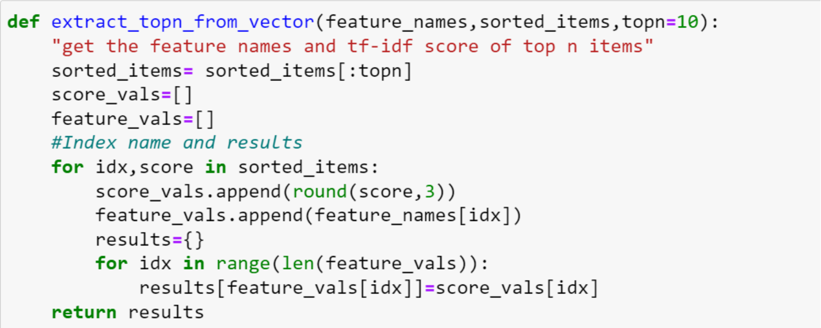
The code is shown below



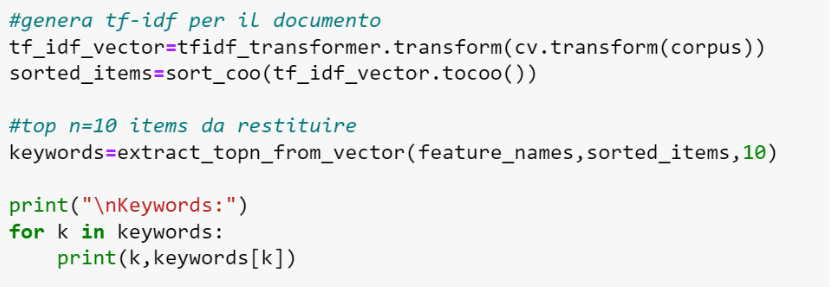
```
#Implementation of TF-IDF
tfidf_transformer=TfidfTransformer(smooth_idf=True,use_idf=True)
tfidf_transformer.fit(X)
#get feature names
feature_names=cv.get_feature_names_out()
```



```
#Ordino i risultati di tf-idf in base al punteggio
def sort_coo(coo_matrix):
    tuples= zip(coo_matrix.col, coo_matrix.data)
    return sorted(tuples,key=lambda x: (x[1],x[0]), reverse=True)
```



```
def extract_topn_from_vector(feature_names,sorted_items,topn=10):
    "get the feature names and tf-idf score of top n items"
    sorted_items= sorted_items[:topn]
    score_vals=[]
    feature_vals=[]
    #Index name and results
    for idx,score in sorted_items:
        score_vals.append(round(score,3))
        feature_vals.append(feature_names[idx])
        results={}
        for idx in range(len(feature_vals)):
            results[feature_vals[idx]]=score_vals[idx]
    return results
```

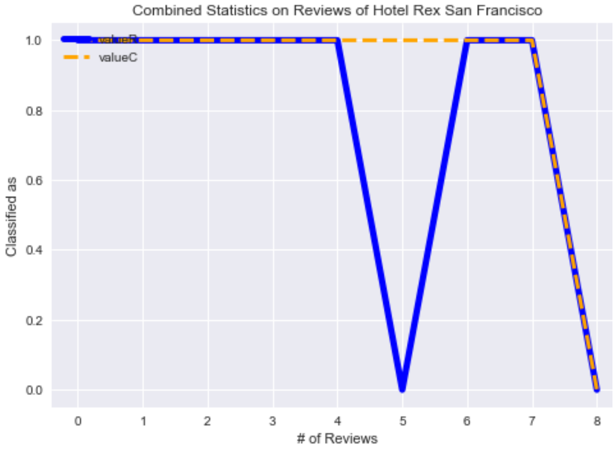


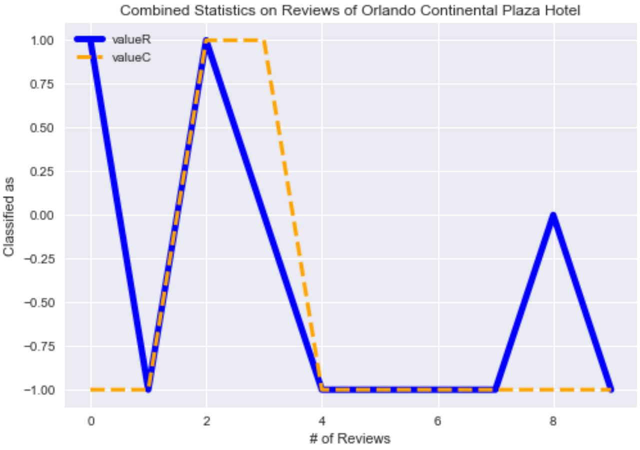
```
#genera tf-idf per il documento
tf_idf_vector=tfidf_transformer.transform(cv.transform(corpus))
sorted_items=sort_coo(tf_idf_vector.tocoo())

#top n=10 items da restituire
keywords=extract_topn_from_vector(feature_names,sorted_items,10)

print("\nKeywords:")
for k in keywords:
    print(k,keywords[k])
```

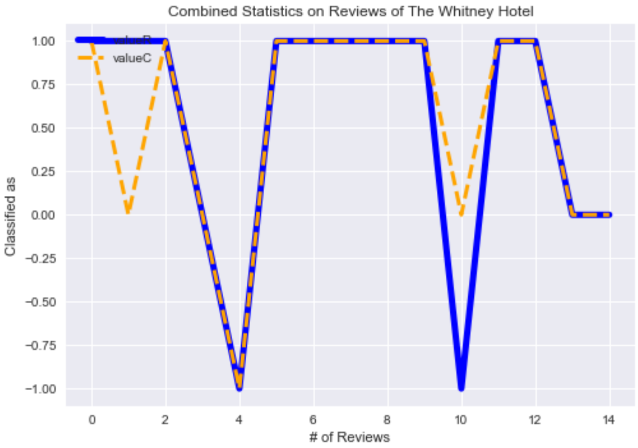
RESULTS DISCUSSION

First Question

Assuming the graph given by the rating (BLUE) as a real trend AND the trend calculated with our classifier (ORANGE).







Analyzing the resulting graphs about the three Hotel, we can see that the results obtained by the automatic classifier don’t differ so much from the people evaluations, **so this software could be used in real life**, obviously it should be refined more, perhaps considering a larger dataset in order to increase the model accuracy.

We compared our software with software currently on the market and it was found that the achievable accuracy is ~ 85%.

Our hypothetical software could be used by companies for various purposes such as:

- Have an overview of the reviews of their products, aggregating the results of different sites;
- Use sentiment analysis to instantly know when the brand is mentioned in a negative way, in order to understand the origin of that data and not ruin the reputation.
- Use this tool to understand in automatic way , the satisfaction of the clients after using customer service of the company
- preserve the online reputation of the company

Or even more for trends analysis:

- See positive and negative mentions volume and reach to see how your reputation changes over time;
- Track reach and learn how many people talked about your new campaign or a press release. To understand which hut has the best performance;
- Track the behavior of the entire sector or of a direct competitor to understand if there is a possibility of improvement and diversification of the product. Perhaps by modifying the software you could also understand the negative and positive aspects mentioned in the review so that you can more accurately extract the data.
- By analyzing the reviews of the company and the sector, it would be possible to create a "Perceptual Map", selecting the attributes that most interest us for our analysis.

From the theory we understood that analyzing the feedback is very important, because it could increase the experience and implement the " learning curve effect" , furthermore the characteristic of Ai companies is to learn much faster than traditional companies and social networks are used by 3.78 billion people, with a growth prospect with thousands of daily review left.

Strategy for our Services:

We imagined having to sell this service, and in this list we will describe the highlights of our strategy:

- we must avoid having a few large buyers, so as not to give them too much "Power of Buyer"
- we must create a service that is difficult to replace
- being a growing sector we should have less aggressive rivals
- We aim for differentiation, therefore with unique features different from competitors. Such as an after-sales team that analyzes customer data once a month.
- being a data analytics company, we could easily do a market segmentation
- our services are mainly offered and b2b, so our strategy will have to be adequate, having experienced customers
- Positioning : the company must be credible, clear and consistent in the advertising of the services offered
- being a service, distribution is simpler, in fact we can have a direct channel.
- For communication it might be ideal to choose online communication. Like for example email marketing and social media, a content site on the topic of AI could also be created, to attract new buyers and industry experts, as Bitdefender did.
- Being a service it can be sold internationally, but in this case you have to be careful of the "price escalation" due to various factors.
- If we have to launch a new service, a good pretotyping strategy could be the "Fake Door", as for example did Airb & b

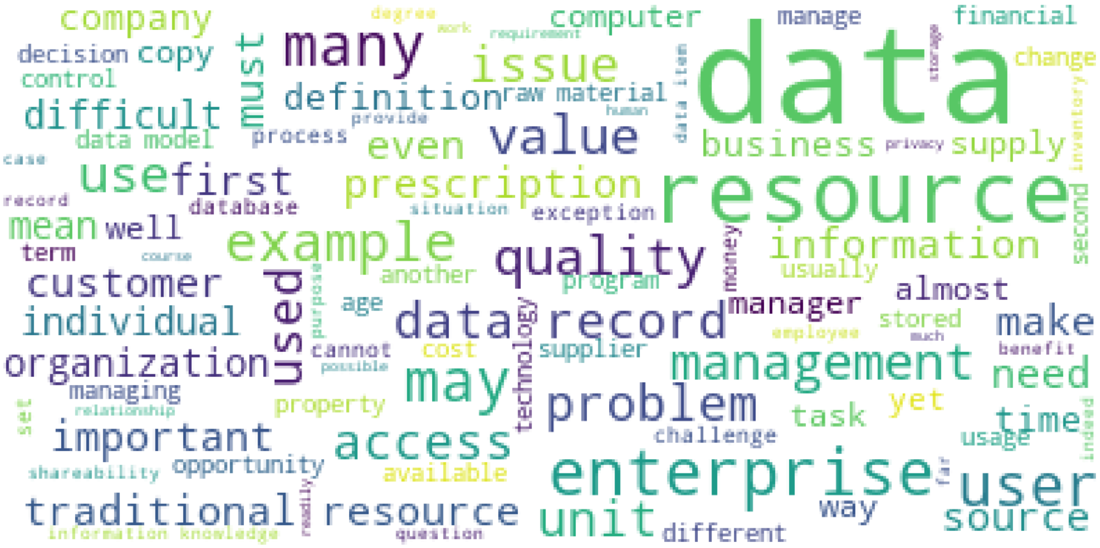
Second Question

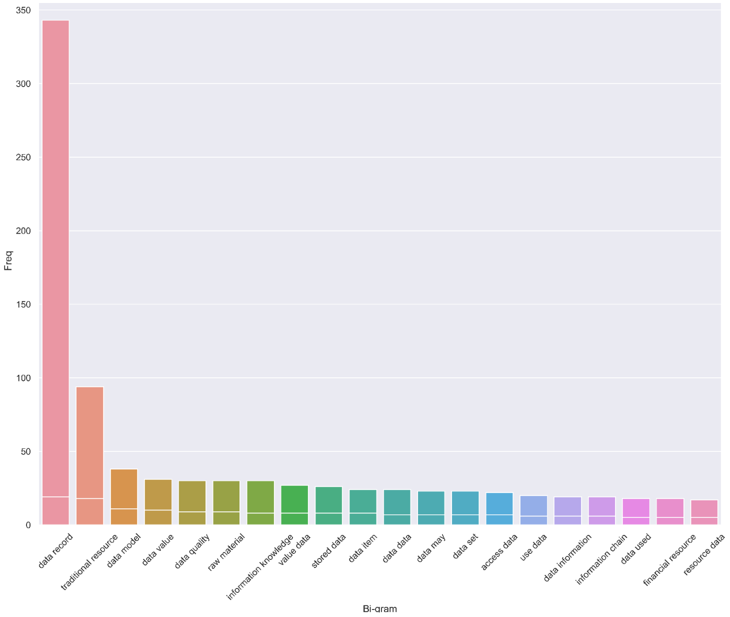
We now proceed to expose the results obtained from the topic finding on the MIT sloan papers.

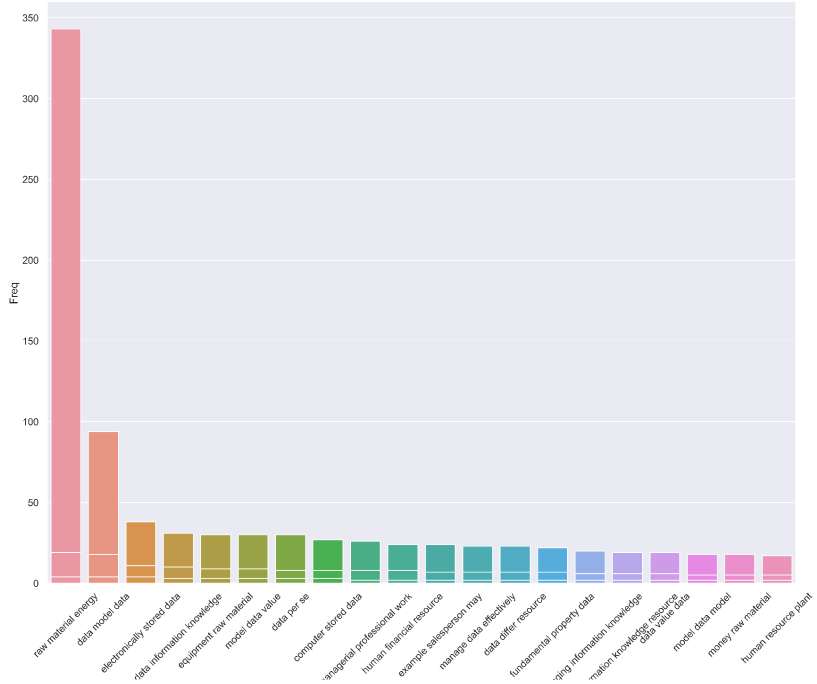
The papers dedicated to the data mining and sentiment analysis field go from 1998 to 2019.

The first paper taken in consideration is “Data as a Resource” dated 15 October 1998.

Just by taking a look at the graphs obtained through the use of the npl process we can observe that in the year 1998 companies already started to interact with electronic data gained from the internet by considering the uses of the user’s information and the problems of storage and maintenance of the data itself comparing them to the traditional ways of data retrieval.



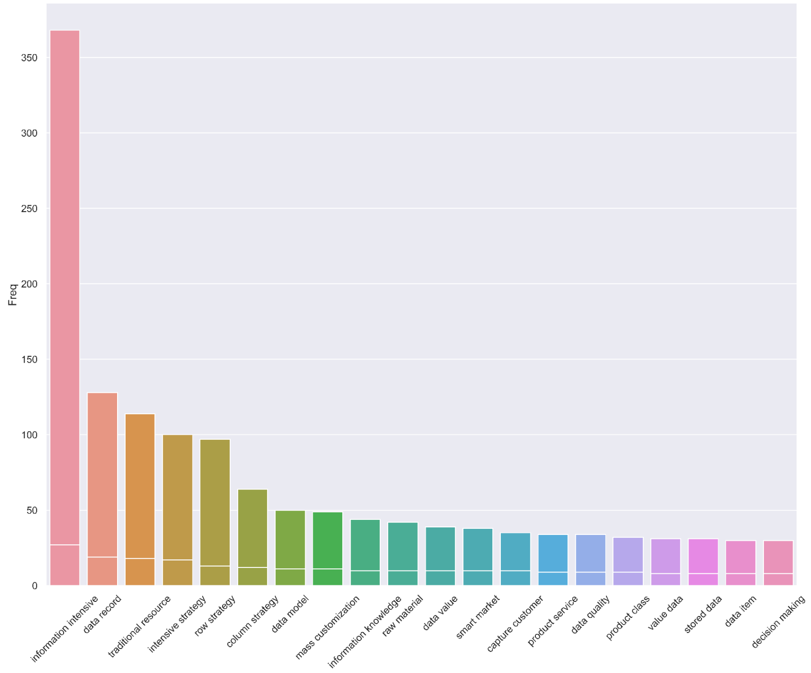


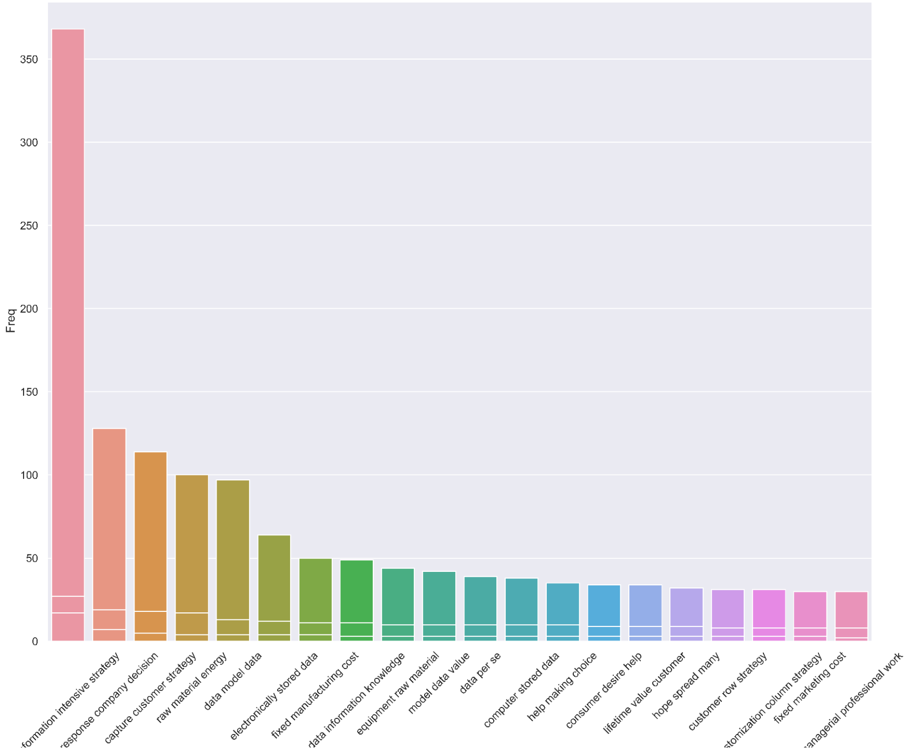


Important topics detected are the ones related to the data treatment and the data model itself. We can understand from this first paper that electronic data was starting to have companies curiosity, seeing in that a possible way to retrieve new information about their customers.

The next document belongs to the year 15 July 1999: “Winning in Smart Markets”.

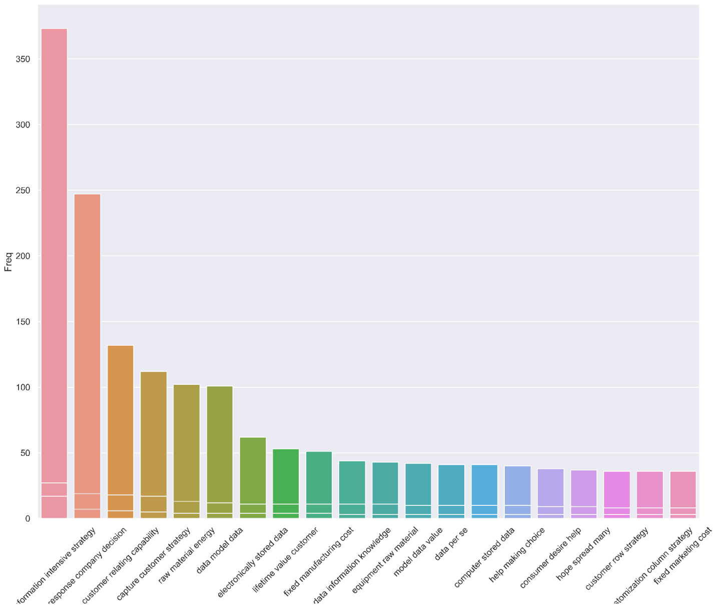
After a year from the first paper we can see that the market companies already started to take advantage of the data analysis seeing in it a great opportunity in order to gain information about the customers preferences and behaviors and use them at they own advantage against the ones who didn’t.



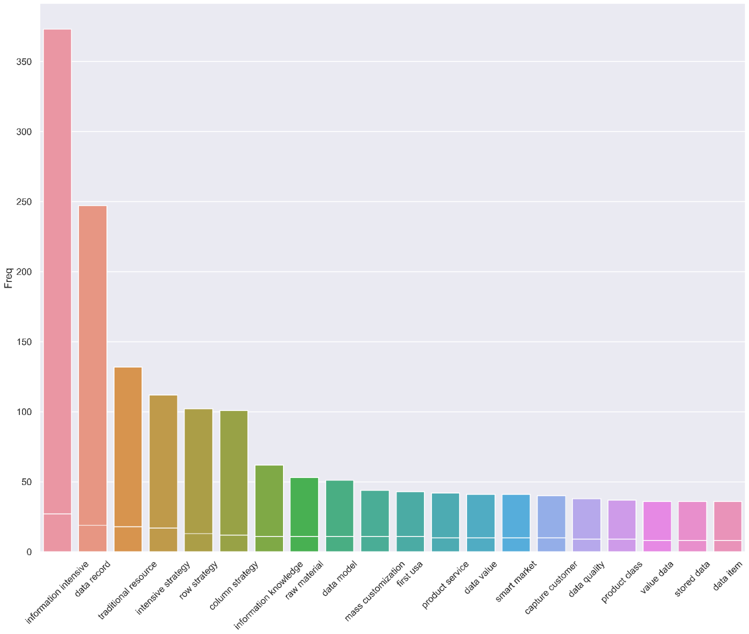


As you can see from the topic detection in the upper graphs “capture customers strategy”, “data information knowledge”, ”help marketing choice”, and “model data value” says a lot about how companies took interest in the digital data field, in particular for marketing problems and solutions and also considering the possible costs of data retrieving.

Moving on, the next paper is dated 15 April 2003, “Creating a Superior Customer-Relating Capability”.



After almost 4 years since the last paper we analyzed, we can see how companies completely integrated the data analysis into their marketing strategy.



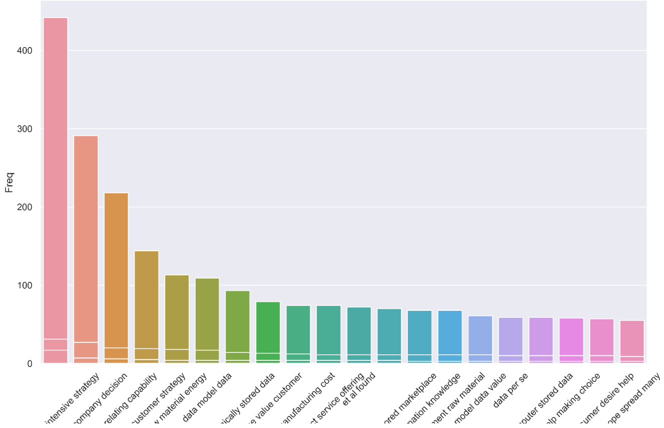
The information retrieved by the graphs are particularly interesting in this case. In this paper are mentioned topics as Smart markets, different types of market strategies related to the customer's data usage, responsive companies decisions, lifetime value customer (which it's started to be also correlated to the feedbacks and the behaviours of the subject not also by analyzing his purchases but also by the digital data gained from a larger view), mass customization and smart markets.

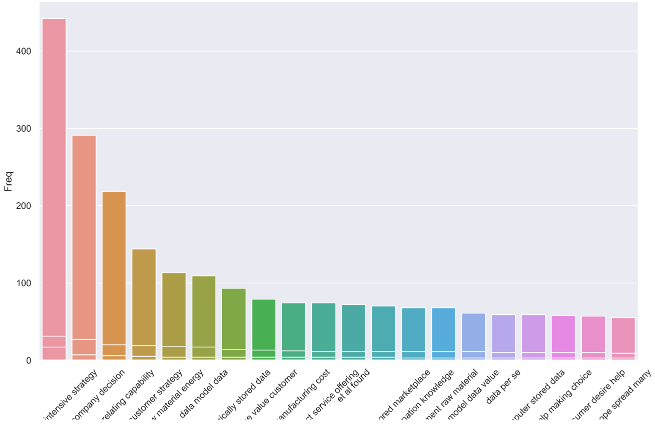
Just by reading the topics retrieved we can already have an idea of how the markets, especially larger companies, started to use the data and sentiment analysis in order to adapt their market strategy and production based on the data acquired. We talk about adaptation of their products based on the customer's taste, which they already did of course, but, thanks to the great speed of gaining and sampling data, it’s has been greatly increased in responsiveness along with the product’s variation based on the people’s feedback.

The market started to adapt very rapidly basing his production decision on the customer’s data, creating new and rapid strategies in order to prevail on other companies.

Beside this in all the papers we can see that the data quality and the cost of data retrieval and analysis is always a concern for the market activity.

The next three papers are from 2005, “E-procurement”, “Evolving from information to insight” dated 15 January 2005 and “Prediction customer choices” dated 01 October 2005.



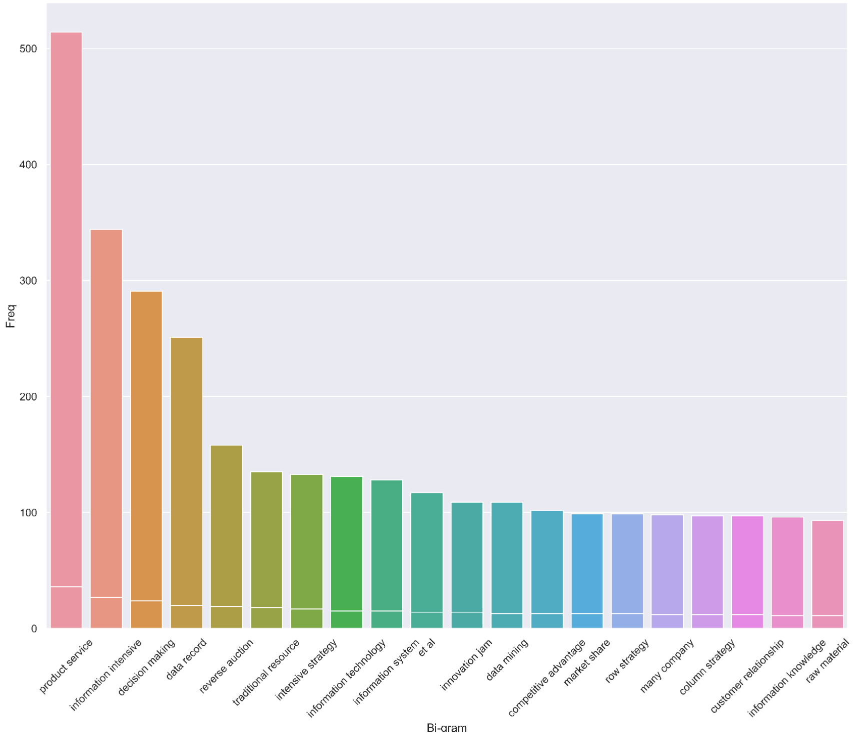


In this papers we can see a consolidation of what observed before. Markets were well integrated with customer’s data treatment and the main concern of the firms verted around the different strategies to use along with the data in their posses. Regarding the bigger companies, intensive strategies prevaled in order to prevail on companies who didn’t take advantage of the data analysis or the ones who couldn’t use it properly.

Still, along with the market strategies companies always kept asking themself which was the best way in order to retrieve and manage customer’s data, both from the cost and analysis techniques points of view.

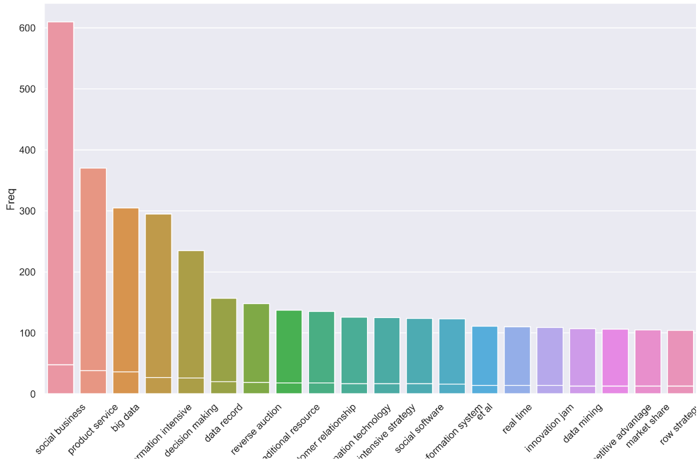
The papers retrieved from 2006 to 2008 didn't give more information than what we've already seen now so we'll move to the one from 2010.
In 2010 the world was going into a new technology innovation, more and more people were starting to buy and use smartphones and that lead to a new and bigger type of data analysis and retrieval for companies in the marketing field.

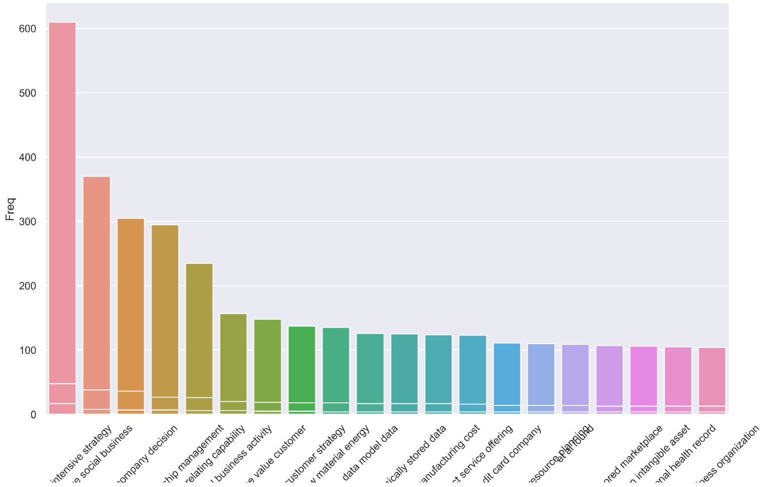
The following graph is taken from the paper dated 26 June 2010 "The IT Audit That Boosts Innovation".



Innovation Jam is the keyword to the new digital world that was about to take life.
In 2010 the social websites were starting to take place and have a huge impact on people's life and also opening new paths for the market to take along with new ways to acquire people and customer's datas.

Let's now consider the paper dated 19 June 2012 "What Managers Really Think About Social Business".



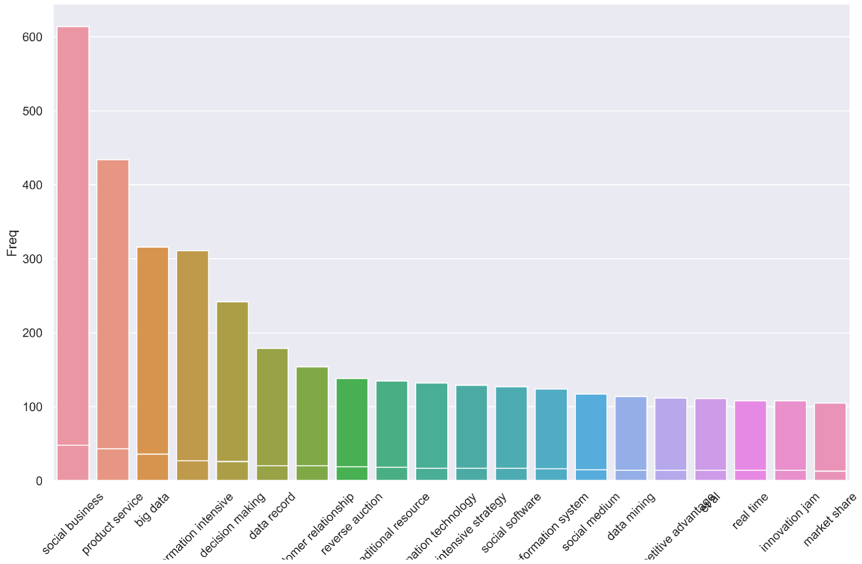


If until now the main topics found were about strategies,data treatment and acquisition here the main one becomes the social business along with intensive strategies.

As we said social networks were starting to have a huge impact on people's life and the business field took interest in the phenomena starting to formulate new marketing actions in order to grow along the new web event that was spreading.

In confirmation of that if we take a look to the topics retrieved by the following paper dated 18 June 2013 "How to Drive Customer Satisfaction" we can see that the social web component has already become a fundamental part in the marketing business field.

Here we talk about "social business", implying that companies not only started to use the social websites like facebook and instagram in order to analyze trends and customers tastes but also to redirect customers own tastes and choices throw their products by precise and direct advertisements strategies.

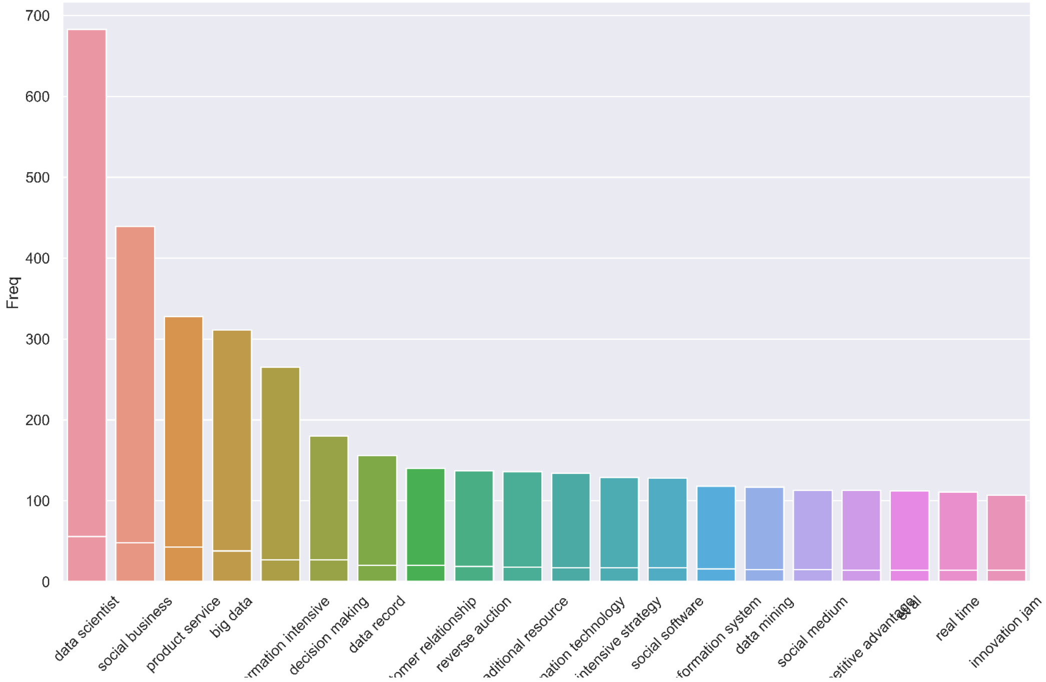


We can already see how in about 12 years the business and marketing topic trends have greatly changed with the evolution of the technology.

Even if we're still talking about customer relationship, business strategies and data analysis, the field in which those elements are considered has changed. Now the internet become fundamental for companies in order to interact with customers and find possible new ones, and social networks became an integrative part of the market analysis as a new field of action.

The following paper is dated 16 September 2014 "Getting Value From Your Data Scientists".

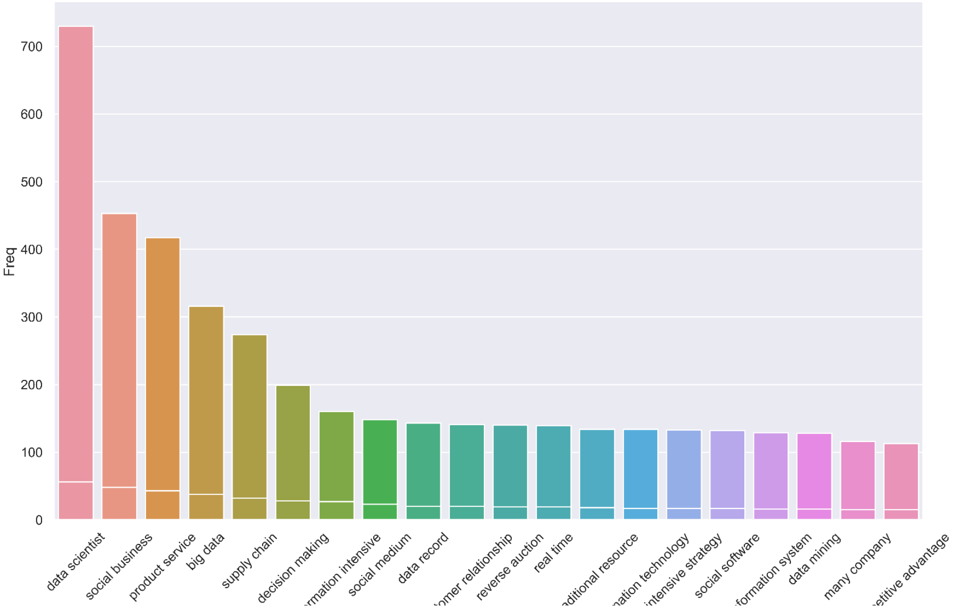
In 2014 and the data available for companies has become immensely larger than before so companies started to search for new figures that were capable to interact and operate with this huge amount of data rapidly and efficiently. They started to look for data scientists.



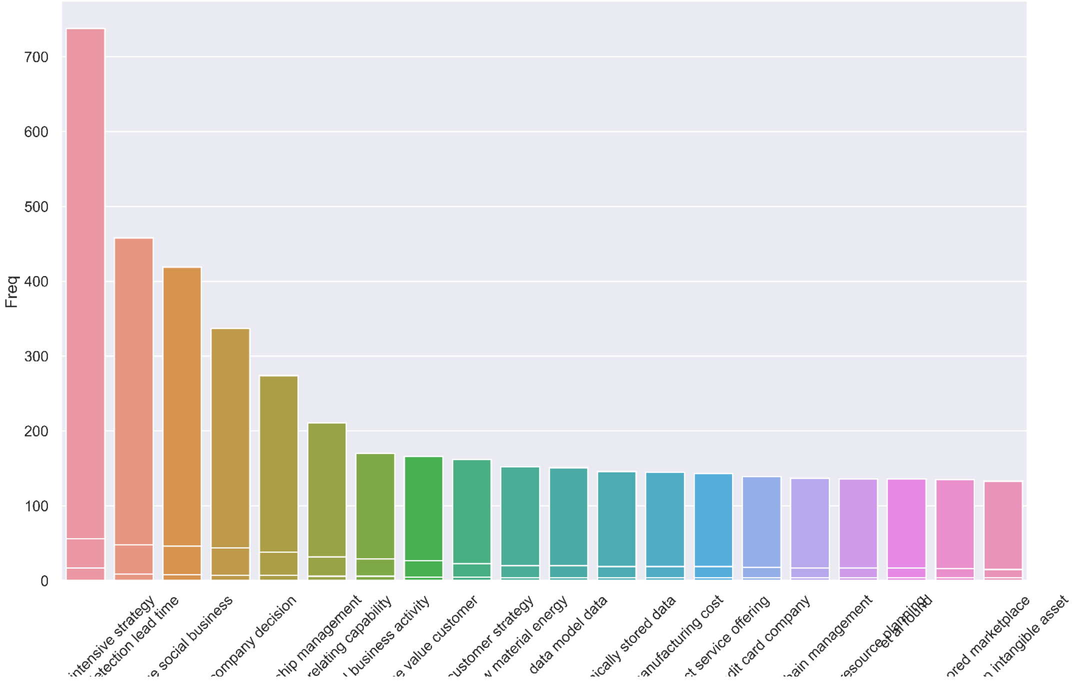
A data scientist is one of the most important new figures of the 21-century. He's capable of working with huge amounts of data by using machine learning algorithms and techniques in order to always retrieve the best quality information from them.

It's easy to understand how having such a figure became imperative for companies who wanted to achieve the best performances and results in a market that was becoming larger and larger, not only physically but also digitally.

If we take a look at the paper dated 15 September 2015 "Preparing for Disruptions Through Early Detection" we can see that after a year the data scientist was a consolidated figure in the market field and thanks to his great capability of analyzing big amount of data companies were starting to being able to take decisions and predictions based on the data analysis instead of manager's intuitions as they did in the past years.



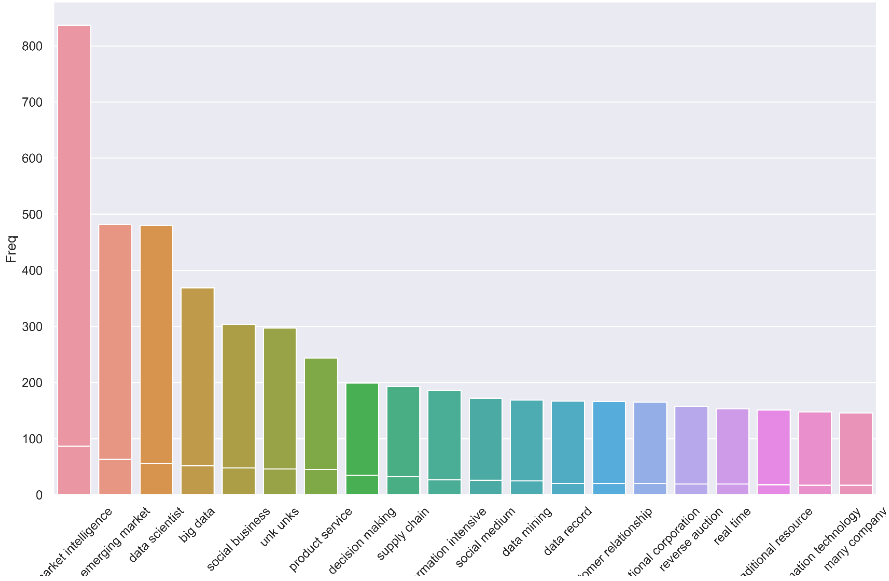
Another important information retrieved from a paper dated 16 March 2015 “Reducing Unwelcome Surprises in Project Management” is the importance of the time needed to retrieve informations. Even working with huge amount of data, it is imperative in this field to be as fast as possible when working with data and this not only confirms the importance that the new figure of the data scientist covered but also how being fast in the data analysis could also save companies from taking or avoiding bad decisions also for prevailing on the concurrence and correct bad management mistakes in a rapid way.



At this point a new breakthrough in the way of how companies operates with big data comes with the usage of the Artificial Intelligence in the business field.

In the paper dated 14 June 2016 “Using Artificial Intelligence to Set Information Free” it’s started to be mentioned the need of the introduction of artificial intelligence techniques in the data analysis and marketing decisions. Even if the topics found are still quite similar to the previous ones we can see that something in the market field was starting to change.

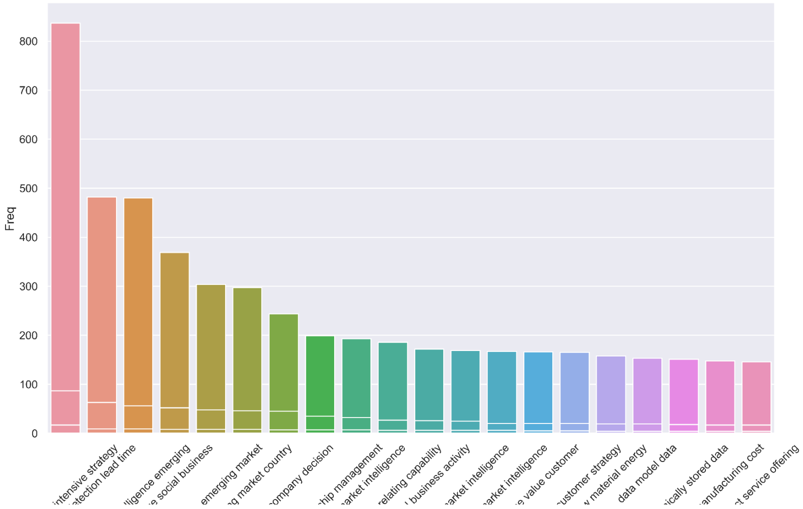
In the paper of 14 November 2016 “Why Big Data Isn’t Enough” we have proofs of what supposed before.



From the topics retrieved we see that A.I. become the main argument for the marketing field. Companies started to adapt their strategies to the growing technologies rebuilding their way of action.

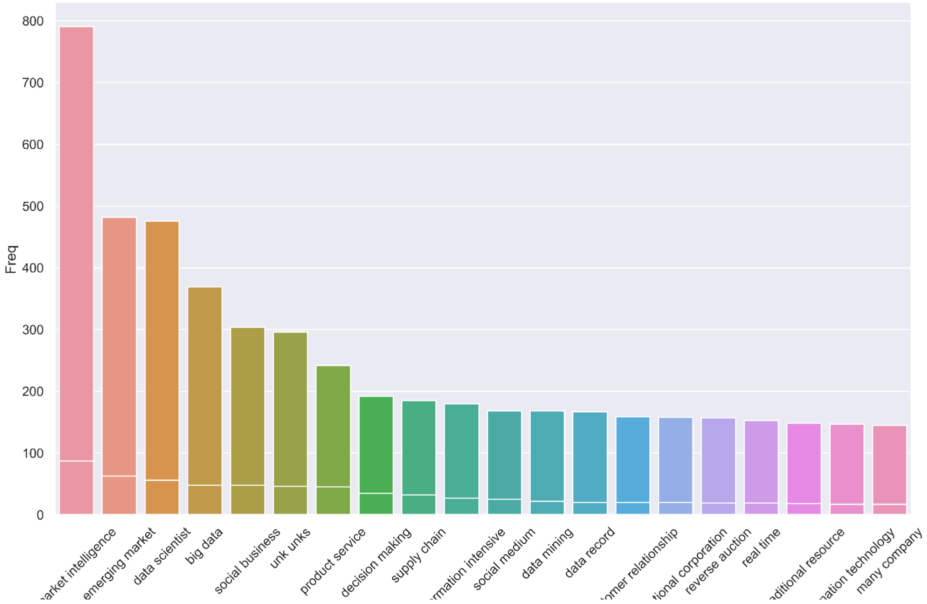
The three main topics let us understand how the whole market nowadays is based on three key figures, the artificial intelligence, data scientist and big data analysis.

The growth of data on the web increases so rapidly that, in order to product a rapid and efficient analysis, always new techniques are required.



The use of data for strategy formulation still remains one of the key elements and problems of the business and marketing field which had to spread and interact with the whole wide world by searching for opportunities also on other countries finding new markets also thanks to the analysis of people’s data and way of thinking through the years.

From the paper dated 5 December 2016 “Mastering the Market Intelligence Challenge” we can see how well consolidated was the usage of artificial intelligence already.



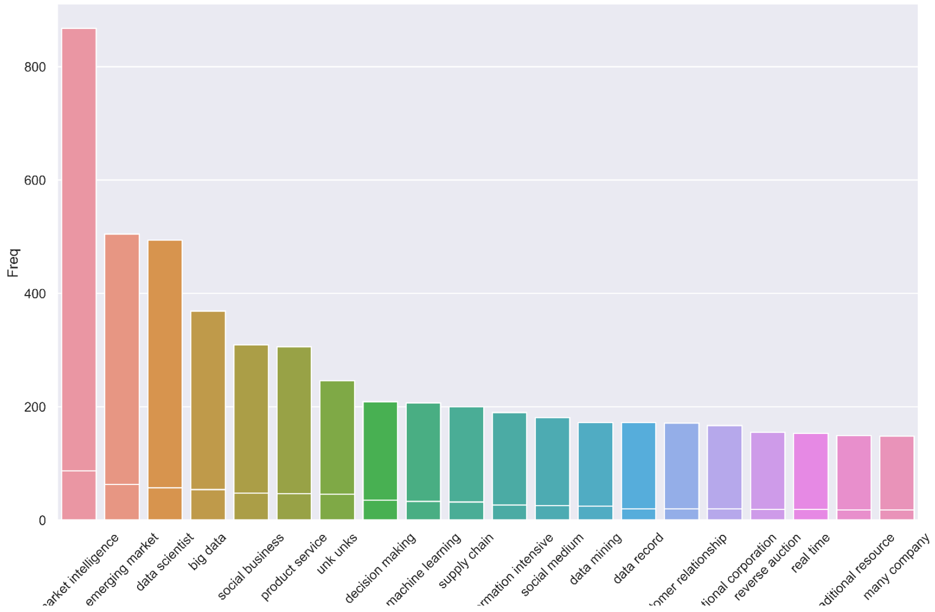
From now on we start to talk about “Market Intelligence” as you can see from the main topic found in the graph.

Market Intelligence was the name given by the process with which a company take and analyze big quantity of data retrieved on many fields (especially from the internet) and use them in order to not only understand people’s feelings,tastes and moving trends but also to use the information retrieved in order to formulate marketing strategies and make predictions about market directions and customer choices.

The capability of a company to analyze, understand and comprehend data in a short time became imperative in order to succeed in the market.

The last three papers are dated 13 March 2017 “Building a More Intelligent Enterprise”, 16 November 2018 “New Ways to Gauge Talent and Potential”, 11 June 2019 “Strategy For and With AI”.

If we compare the graph obtained from the last paper with the last one showed dated 2016 we can see that the topic in these years remained the same.



Artificial intelligence, data scientist, big data and social business are still the main features of the business market nowadays.

From the analysis of the papers we had a vision of how much the market evolved and grown in less than 20 years with the advance of technology.

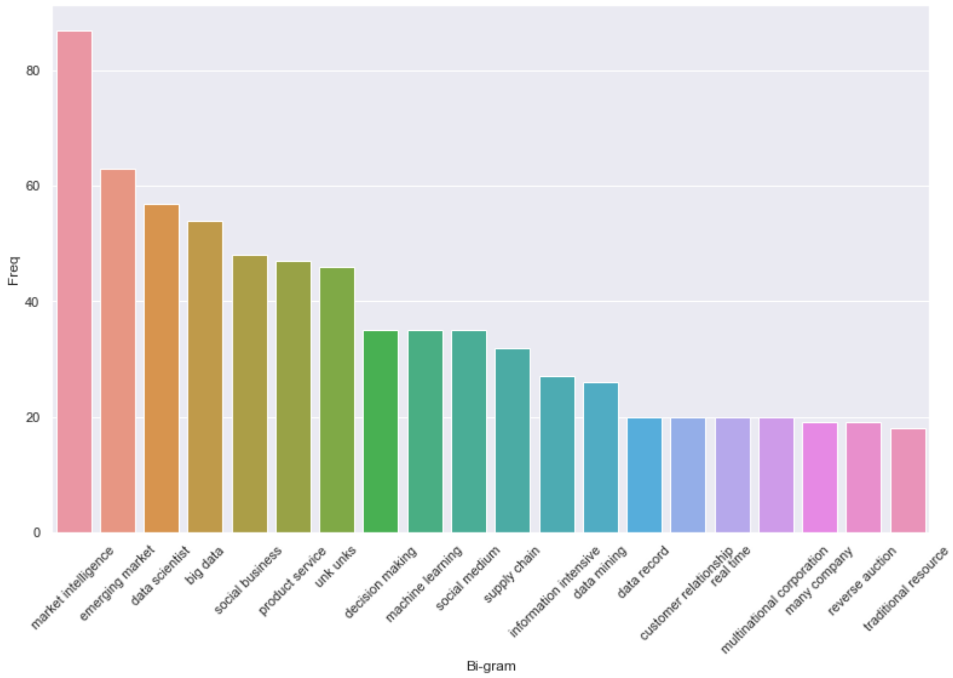
Starting from strategies based only on managers capabilities and intuitions and production needs the market changed basing it's way of acting to the customer's needs and sentiments through time.

Data acquisition has always been a fundamental part for companies in order to make strategies and decisions but the way on which they are acquired changed with the years.

Starting from informations gained on the field, with the growth of the web and the advent of social medias, new ways of data requirement, analysis and usage were born.New figures in the business field started to have a main role and the need of the machine learning and artificial intelligence usage and techniques became essentials for companies.

The digital evolution signed a new step in the behaviour of companies with their customers which didn't have to adapt their tastes to what the market was offering anymore but, instead, their own tastes become a fundamental part of the production decisions for companies.

If we check the graphs obtained from analyzing all the papers together as one we can see that the core of the market strategies is the usage of the AI and machine learning techniques. The data scientist became a fundamental role for business decisions and strategy adoption along with the big data analysis. Social networks are considered the main field on which a company can interact with customers and create new ones. The internet is now the principal field on which companies uses advertisement sell products and interact with people.



In conclusion, topic analysis let us understand how and how much the business and market field transformed and adapted itself during those years.

We can understand by this study how much this two subjects are strictly related with the technology evolution and how is important for a company to always keep up with the world’s digital evolution both from a technological point of view and from the customer’s perspective.

Sentiment Analysis plays in this case one of the key roles in the social data analysis. Throw this technique a company will always be able to know the actual perception of people about a product or the company itself. Based on these feedback firms can improve their image along with their products in order to gain people's preference.