Example 1

The corpus at hand, Quotebank, is a collection of 12 years of quotes from the American press, which were extracted using Quootstrap and attributed to their speaker using Quobert.

In my opinion, the American press is a perfect reflection of American society, and carries within it the questions and answers of the future of the country. That's how I was inspired to explore both political and economical issues. I would like to study and try to answer these three problems that interest me personally.

The presidential election: Who will be the next president of the united States ? We could use the quotations of the former US presidents the year before their election, to predict how a future presidents speaks publicly. We could see the citations of the presidents for the elections of 2008, 2012, 2026 and 2020, and try to find what are the the words that allow them to be elected. Do celebrities have any influence on the results and popularity of political figures?

Climate change: After the Ipcc, the Intergovernmental Panel on Climate Change, a few weeks ago, today the United in Science 2021 report warns us of the global climate situation. According to the scientists, we are "way behind" on the objectives of the Paris Agreement. Is climate change a topic that politicians only talk about during voting periods ? What are the periods when the subject is most highlighted, and at most at the heart of the debates ? Do politicians all have the same words and ideas on the subject, or do they make the dread of bringing a personal reflection on the subject ? We could evaluate the occurrences of the total citations about « climate change » by having a list of climate-change-related words , and look for some tendencies.

The price of cryptocurrencies: What has been the evolution of quotes about cryptocurrencies since 2008 (date of creation of the bitcoin) ? Is there a correlation between quotes on cryptocurrencies in newspapers, and the price of these assets ? We could focus on periods of crises or large increases in the price of cryptocurrencies, to see if the publications have an effect on the price. And then search if political or governmental speakers have more influence than specialist people on the subject. We could compare the times-series of the number of occurrences of a positive crypto-related word, with the time-series of the price of the Bitcoin (BTC), as the Bitcoin is the currency that influences all of the others, and search for a correlation.

Example 2

Idea 1: Brexit deal: who wins and who loses?
On 31 January 2020, following the official exit of the UK from the European Union, Michel Barnier stated: "There is no winner in the Brexit, it is a lose-lose situation". However, the reality is more complex, and it is likely that some sectors are benefiting from the crisis. The aim of this study would therefore be to analyse all the quotes related to the Brexit, in order to identify which sectors are benefiting from the deal, and which are suffering from it. For example, British nationalism could emerge stronger from the Brexit, while the economy will probably be weakened. For each sector of society analysed, the methodology would be the same: the idea would be to use a sub-field of natural language processing called sentiment analysis, in order to quantify whether the sentiment is rather positive or negative for the quotes related to this sector (binary sentiment classification). As this is a complex classification task, which depends a lot on

the application domain, it could also be interesting to use different classification tools and then aggregate the results to decide on the sentiment intensity, with a vote based mechanism for example.

Idea 2: China in the eyes of the West
The emergence of China as a major economic power has been closely followed by the international press for the last 20 years. In the West, China is shaking up the way people think and there are different views on how to deal with the new power. The objective of this study would be to analyse the evolution of the Western view of China's development through the lexical field used (admiration, fear, intrigue, obsession, etc). In other words, the idea would be to carry out a sentiment analysis to follow the evolution of different emotions towards China over time. For example, one could show the obsession with China in the American discourse since Trump's mandate. As with the first idea, it might be wise to use different classification tools, for example by combining the use of lexicons with machine learning algorithms.

Idea 3: Between the older and the younger, who is more articulate?
A common stereotype is that older people, because of their experience, are better at speaking and writing than younger people. This study aims to determine, by means of a data driven approach, whether there is a real basis for this fact, or whether it is just a cliché. The assessment of the level of language could be based on various criteria including the range of vocabulary and the use of logical connectors. It could also be interesting to reveal new linguistic usages that have emerged in the last 20 years. One of the challenges of this project is to access the age of the speakers at the time of the quote. For example, we could use the speaker's Wikipedia ID to determine his or her date of birth and deduce the targeted age.

Example 3

How has “wokeness” progressed with time? With more people being given the right to speak up, we find out about different struggles and points of view in life, and thus we become more woke (i.e., aware of society’s injustices). It would be interesting to see whether, with time, people have become more sensitive to social injustices and if it reflects in their speech. To pursue such a study, some sentiment analysis on quotes about social topics, such as black lives matter, LGBTQIA+ rights, and sexism, needs to be done. In addition, a neural network that could detect racism/sexism in quotes could be trained. However, since the quotes we have are not labeled as being or not being offensive, we could train the neural network on another labeled dataset. Then, we could use the trained neural network to generate labels for the Quotebank dataset. Instead, we could also use unsupervised learning. Using these tools, we can observe the increase/decrease of insensitive quotes made in the past versus now.

Is your voice that unique? We say that each person has a unique voice; however, is it possible to predict your background based on something you have said? This project would focus on the study of the influence of a person’s profile on their speech. A person’s background, education, age, etc., can shape one’s manner of speaking. Therefore, a link between demographic factors and the style of speech might exist. Thus, given a quote, can we predict the speakers’ age and race, for instance? To learn such a model, we would require a dataset that includes information about the speakers. One such dataset is Wikidata. Thus, combining the two datasets, we could learn a model that predicts demographics based on speech.

Is it possible to learn the style of some people and then generate quotes that could have been said by them? Currently, the model learned for word prediction on phones is personalized for each user. The model can capture a person’s writing style and suggest endings for an incomplete sentence. Can we push it further and have models that can generate from nothing a complete sentence? Focusing on a

subset of speakers, we could train a generative adversarial network (GAN) to learn their style of speech and to generate quotes. The model could also be trained to predict a person's stance on a given topic. Then, a follow-up question would be, are certain speakers easier to learn? And if we were to generate a fake quote, how likely would it be believed? If the current dataset is not enough to conduct such a study, we could rely on tweets sent out by the speakers in the dataset with a Twitter account.

Example 4

Idea 1: Bitcoin's 2017 Boom, Manipulated bubble? In the last five years, the cryptocurrency market had significantly gained in sight. Not only has it attracted institutional interest, but also retail demand. In fact, the year 2017 was unprecedented in the history of Bitcoin: Over 12 months, the prise rose about 20-fold before losing 65% of its value some months after. In this context, many individuals invested their life-long money savings as a natural reaction of the fear of missing out the century's opportunity. This FOMO was partially fuelled by notorious entities publicly projecting their prices expectations. Given that market manipulation is prohibited by all the financial market regulators. Light should be on this matter. Several big investors are involved in the industry, such as Jamie Dimon, Elon Musk, Mark Cuban, and Tony Robbins. Using QUOTEBANK it is possible to investigate the correlation between bitcoin prices and how optimistic/pessimistic investors' speeches were. It is possible to extract all the quotations made by these public figures in time interval centred on 2017 and filter those with the mention of "Bitcoin" and interpolate them into 2 clusters: Bullish ones and Bearish ones and match each cluster with the prices of Bitcoin for that time.

Idea 2: The evolution of presidential premises before and after 2016 elections in the United States. In political sciences, the election speech is considered one of the most important components in the process of motivating and convincing individuals to vote. During those speeches, each candidate clearly draws an image of their ambition to tackle the three main issues of the country: Healthcare, Immigration and Education. However, it is well known that the focus shifts towards more actual topics as Economical improvement and External relationships once elections end. Using QUOTEBANK, it is possible to gain understanding about the main issues that the former president Donald Trump promised to tackle and what he really tackled during his presidency. The idea is to extract all quotations said by the former president and filter them by date: Before his presidency and during his presidency. For each of these clusters, find the n-most frequently mentioned topics and draw a comparison between pre and post presidency. Results can be confirmed by adding new data: typically, the evolution of the US HDI (Human Development Index), GDP (Gross domestic product), Immigration flows and healthcare coverage.

Idea 3: Brexit: Were the results predictable? In 2016, the European union has been shaken by the vote of the United Kingdom to leave. The referendum results were extremely close: 52% vs 48%. The full effects of the Brexit have taken place only four years after. The debate had been around well prior to the referendum. QUOTEBANK can be used in this case to evaluate the predictability of the referendum's results by extracting all the quotes containing the word "Brexit" or "United Kingdom leaving" and group them into 3 categories: "Obviously leaving", "Obviously staying" and "Neutral" and see how close were the two obvious groups to the real results.

Example 5

The following ideas are proposed by me. Note that we can come up with a project which involves all 3 ideas.

a. Categorization of News: We notice that for each quote we are given URLs for the comment. Each quote has at least one URL for their NYT article. It is easy to see that for NYT URLs, we can easily find categories of different quotess by doing some pre-processing on the NYT URL string. For example, one of the URLs is https://www.nytimes.com/2019/04/17/realestate/house-hunting-in-hong-kong.html?partner=rss&emc=rss . The URL follows a similar pattern in most cases with the name of the website followed by the date and then the category of the quote. Once a category is assigned to each quote, we can do many interesting analyses such as what categories are the most popular with different news providers and speakers. If the number of occurences can be taken as an index for popularity, we can also see which categories of news are more popular (i.e shared by other news articles) and which are not popular.

b. Graphical analysis of news vendors and speakers: Since there are many quotes with multiple URLs, we can plot different graphs which show how interconnected news vendors are between themselves. This interconnectedness will be defined by the amount of times 2 different news vendors share the same quote. We can also have an interconnection between different speakers and news vendors. It could also be interesting to split up NYT into different categories like NYT-Climate, NYT-Real Estate, etc as defined in the previous section and then drawing these graphs. We should expect a clustering of nodes around these category nodes with weak connections between them.

c. Bias in Media: In this project, we propose that we train a sentiment classifier which can predict the sentiment of quote as positive or negative. For this, we'll most likely need an external dataset which helps us train our sentiment classifier. Once we do have this, we can understand for different speakers, what has their overall sentiment for different categories been. Has their and overall opinion about a category changed over the years. We can also study what makes news articles viral. Since we have the number of occurrences of each quote and also their category now, we can understand for which categories and sentiments are news articles more viral than the others.