Tina Hegelund Hyldahl Martin - HSR607
Mikkel Birk Nielsen - FXK415
Kasper Senika Larsen - PHM536
Christian Runge Hansen - QKT122

Københavns Universitet
Open Data Science
Informationsvidenskab
Vinter 19/20

Portfolio 2

Key Findings and Reflections

For this assignment we decided to work with TheGuardian OpenApi. This decision was based on the fact that TheGuardian OpenApi was more similar to the 20newsgroup dataset, we had worked with in class.
It also seemed to be presented more orderly than the transcripts of danish news broadcasts, which would give more options for processing the data. With the lack of punctuation in the transcripts of danish news broadcasts, processing would be difficult, as we would not be able to define end of sentence markers.
Furthermore TheGuardian OpenApi would give a wider range of topics to research, which would give more freedom in choosing a research question.

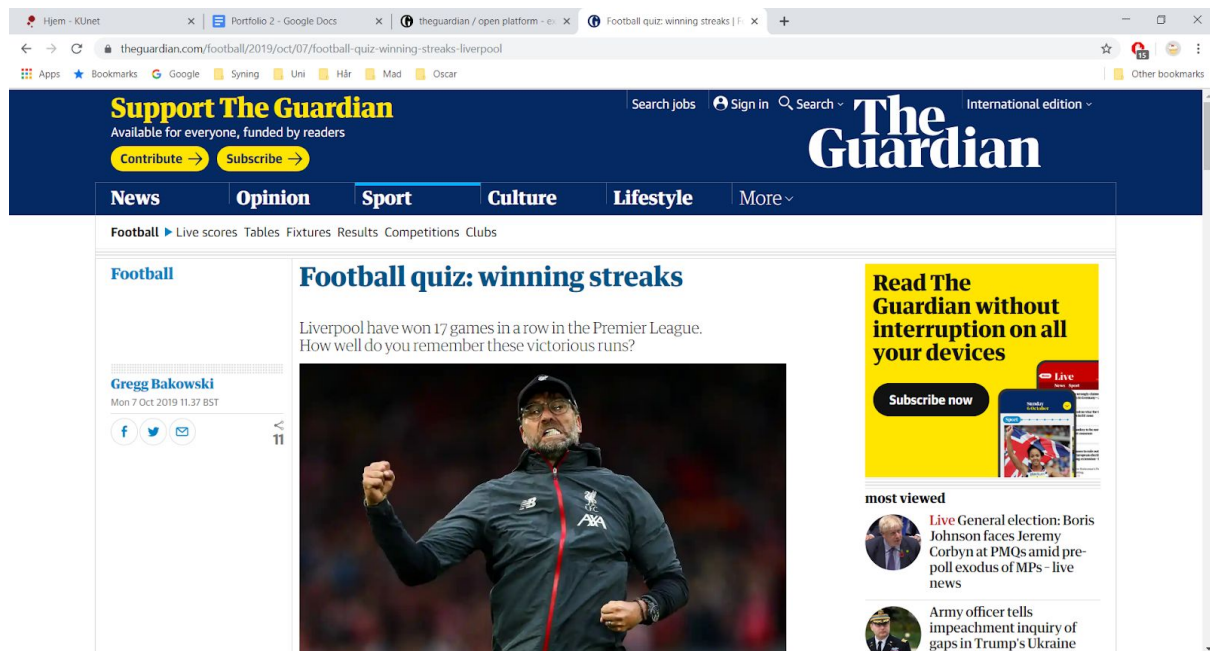When exploring the dataset on The Guardian Open Platform (https://open-platform.theguardian.com/explore/), we noticed that all articles were listed by multiple different categories:

```
▼ 0:  {} 11 keys
    id: "football/2019/oct/07/football-quiz-winning-streaks-liverpool"
    type: "article"
    sectionId: "football"
    sectionName: "Football"
    webPublicationDate: "2019-10-07T10:37:20Z"
    webTitle: "Football quiz: winning streaks"
    webUrl: "https://www.theguardian.com/football/2019/oct/07/football-quiz-winning-streaks-liverpool"
    apiUrl: "https://content.guardianapis.com/football/2019/oct/07/football-quiz-winning-streaks-liverpool"
    isHosted: false
    pillarId: "pillar/sport"
    pillarName: "Sport"
```

From these categories we chose to work with "id" and "sectionId" when importing the dataset.
We chose to work with "sectionId" over the very similar "sectionName" as it had the name of the section listed in small letters, which is easier to work with in python, as python differentiates between capital letters and small letters.
We chose to work with the sections to be able to do a broad categorization of all the topics for further analysis.

The overview provided by The Guardian Open Platform helped us understand what the different categories were used to show, as well as where they would show when viewing the articles in a browser:

Tina Hegelund Hyldahl Martin - HSR607
Mikkel Birk Nielsen - FXK415
Kasper Senika Larsen - PHM536
Christian Runge Hansen - QKT122

Københavns Universitet
Open Data Science
Informationsvidenskab
Vinter 19/20

Data processing

During the pre-processing phase we noticed that the total word counts using the str.split and the word count tokenizer did not show a big difference when compared to each other. We assume that this is because the data in The Guardian collection is well-edited news materiel.

We used random samples of the unique words we got when using str.split. This was done to get an idea of what the unique words look like. Here we noticed that some words were not typical words e.g. numbers: '£22.99' or abbreviations: 'JCRA' or that the words had punctuation marks around them and therefore appeared as a different word than the same word without them: look" vs. look. Therefore we used the nltk tokenizer to weed out all of these unique words that were not actual words. The nltk tokenizer automatically converts all capital letters to small letters, to make sure all unique words only appear once.

Compared to our previous work with the 20newsgroup dataset we also noticed that among our sampled unique words there were fewer spelling mistakes. We assume that this is due to the fact, that the data in The Guardian collection is published news materiel, that is edited before publication.

It is interesting to notice, that when removing the stopwords from the full dataset with 30 million words, almost half of all the words are removed. But when working with the unique word count, applying the stopwords only removes 147 unique words.

When looking into the most used words for the complete dataset, all the top words: 'said', 'one', 'would', 'people', 'also', 'new', 'time', 'like', 'first', 'says' - are very common words that are likely to appear in the majority of all articles.

After this count we decided to add more stopwords: 'said', 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight', 'nine', 'ten', 'would', 'could', 'says', 'also'. This gave a new list of the most used words for the complete dataset: 'people', 'new', 'time', 'us', 'like', 'first', 'last', 'years', 'government', 'back'. We also did the list of top ten most used words for the "Football" and "Sport" sections twice, before and after adding more stopwords.

Some of the top words from the list of most used words for the complete dataset, were also among the most used words for the "Football" and "Sport" sections. The words on the list of most used words for these two sections, that were not on the list of most used words for the complete dataset, give a good indication that they are related to football and sports: 'one', 'first', 'game', 'back', 'two', 'time', 'last', 'england', 'ball', 'said'.

Words like 'game', 'england', and 'ball' are commonly used to describe sports and football. The fact that 'england' is among the top words also gives a clue that England's own national teams are in focus.

In the second run of top ten most used words for the "Football" and "Sport" sections with added stopwords, it is even more clear that the words are related to sport: 'first', 'game', 'back', 'time', 'last', 'england', 'ball', 'team', 'world', 'min'. Here the words 'team', 'world', and 'min' are added to the list.

Research question and restrictions

We have decided to focus on the "Football" and "Sports" sections. We have chosen these sections because they are both larger sections with many articles, but with comparable and similar topics.

We have decided to use "With which topics was Tottenham (football team) mentioned?" as our research question, as it was recommended to use a named entity in order to not overly restrict the number of topics in our subset. It is a quite specific named entity as we felt like anything less specific like e.g. "ball" or "football" would be too non-specific to look for within the "Football" and "Sports" sections.

Before making a final decision, we also experimented with different queries such as: "With which topics was Brexit discussed?" But we are all tired of hearing about Brexit. We also talked about how interesting it was, that Brexit is a part of the politics section, rather than it being in its own separate section, considering how this topic dominates the politics sections of many news providers at the moment.

We also discussed how looking into the musician "A$AP Rocky" could be difficult, as his name might be caught in the nltk filter due to the use of $ in his name.

In the following of our research question: With which topics was Tottenham (football team) mentioned? We expect to find the topics: Tottenham, hotspurs, spurs, stadium, hotspur, new, and home. This is because we know that the players on the Tottenham football team are referred to as the "hotspurs" or just the "spurs", and that they had a new stadium built in the spring.
We have chosen our query due to the expectation of these terms being used by the journalists; e.g. we expect the nickname of the football team to be used to refer to the team.

Furthermore, we expect to find topics such as: NFL, football, new stadium, transfers, (Christian) Eriksen, Champions League, and National Team.

With the top 5 most relevant articles that we found through our query, we try to answer our research question: With which topics was Tottenham (football team) mentioned?

1

https://www.theguardian.com/football/2019/jul/02/tanguy-ndombele-tottenham-medical-lyon-roma-toby-alderweireld

This article is about the transfer of Tanguy Ndombele to Tottenham, whilst also discussing other possible incoming and outgoing transfers.

2

https://www.theguardian.com/football/blog/2019/jun/02/mauricio-pochettino-harry-kane-tottenham-champions-league

This article is focused on the manager of Tottenham and reflects over future choices following their champions league final defeat.

3

https://www.theguardian.com/football/blog/2019/may/29/why-i-gave-away-my-champions-league-final-ticket-to-watch-it-with-my-dad

This article is about how a man trades away one expensive ticket for the final match of champions league, so he can watch it with his dad instead.

4

https://www.theguardian.com/football/live/2019/sep/08/chelsea-v-tottenham-womens-super-league-live

It is interesting to notice that this article is about women's football - still the Tottenham spurs - but not the men's team like we expected.

5

https://www.theguardian.com/football/2019/apr/30/tottenham-hotspur-ajax-champions-league-semi-final-first-leg-match-report
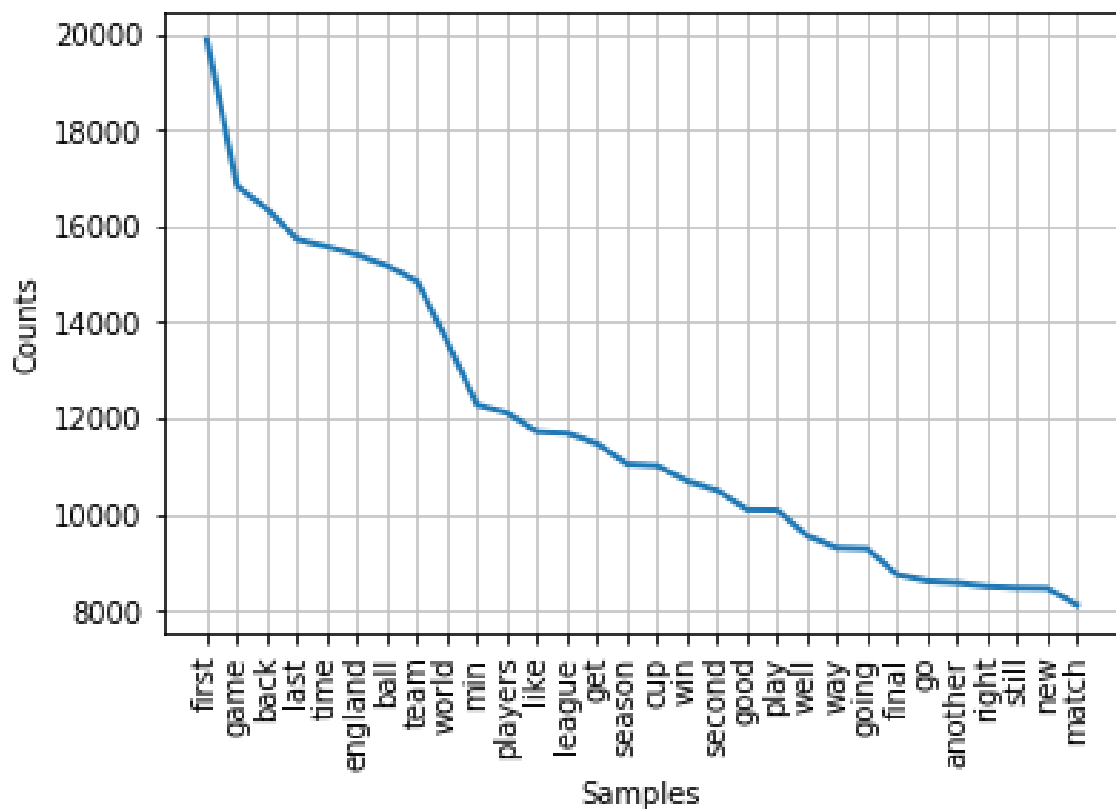
This article is a match summary following a champions league match between Tottenham and Ajax.
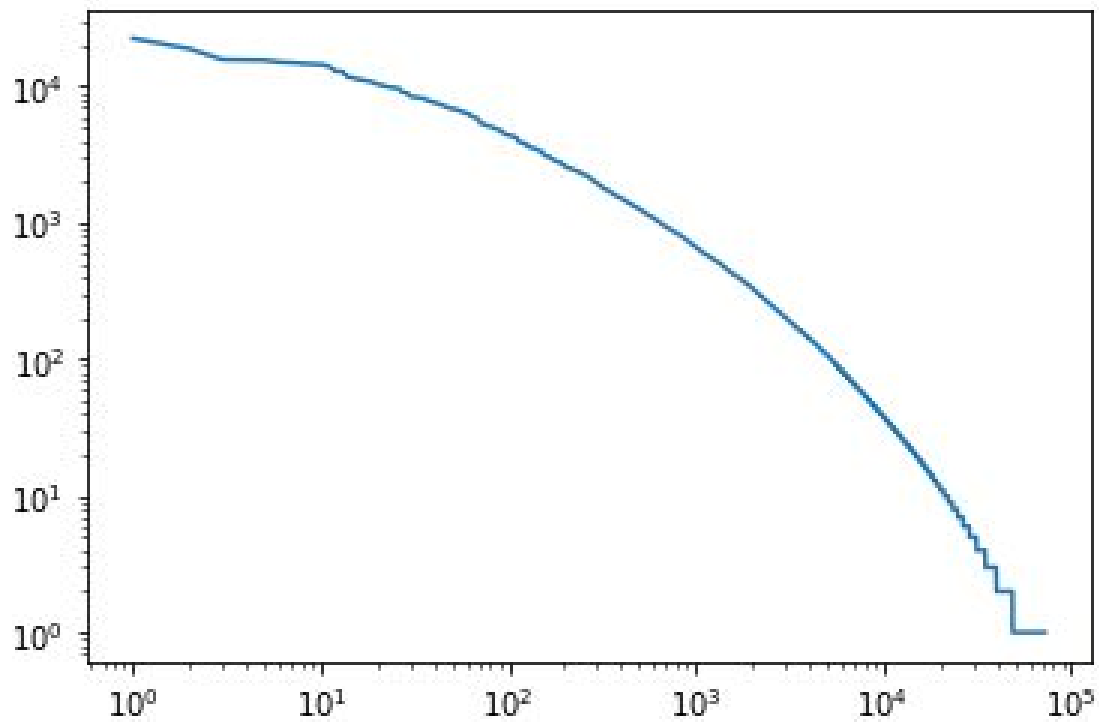
Tina Hegelund Hyldahl Martin - HSR607
Mikkel Birk Nielsen - FXK415
Kasper Senika Larsen - PHM536
Christian Runge Hansen - QKT122

Model and visualization of topics

Query terms presented in a tf-idf matrix:

query term	count	idf
tottenham	1975	4.775314
hotspurs	2	10.600331
hotspur	312	6.328305
spurs	2310	5.153593
stadium	2795	4.323687
new	61406	1.614886
home	24605	2.249586

Frequency distribution of the top 30 most used words in the football & sports subset:

Tina Hegelund Hyldahl Martin - HSR607
Mikkel Birk Nielsen - FXK415
Kasper Senika Larsen - PHM536
Christian Runge Hansen - QKT122

Zipf's law graph - showing that in this dataset the frequency of words is inversely proportional to its place in the table:

Tina Hegelund Hyldahl Martin - HSR607
Mikkel Birk Nielsen - FXK415
Kasper Senika Larsen - PHM536
Christian Runge Hansen - QKT122

WordClouds for the topics with the highest weight from 6 random documents:
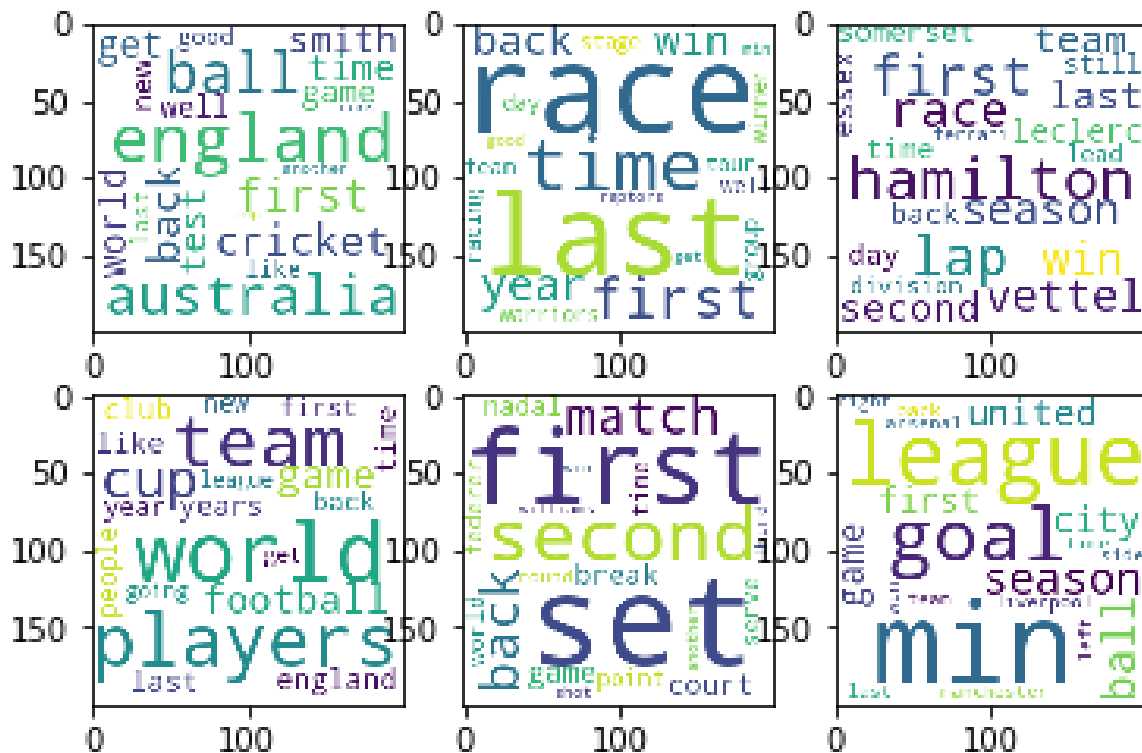
Topic 0:

In the WordCloud for the first topic, among the words with the highest weight we find "england" and "australia". Based on that we can conclude that there's talk about the national teams and not the leagues. The word with the sixth highest weight is "cricket" and that sums up which sport there is talked about and with the aforementioned national teams, what they are playing. Another thing to notice is that "test" and "game" have a very similar weights, that can indicate that this is a test game between England and Australia, but nothing that can be confirmed for sure.

Topic 1:

Two of the most indicating words in topic 1 is "warriors" and "raptors". These are both basketball teams in the NBA, and therefore they could indicate that the article is about basketball. There are also some words there are kind of hard to connect with basketball, such as "race", "tour" and "stage". Therefore it is hard to say if it is basketball the article is about but other words such as "win" and "min" could also lead against its a basketball article.

Topic 2:

The topics for this article are clearly indicating what type of sport the article is about. The word with the highest weight is "hamilton" and two other words with a respectively lower weight is "vettel" and "leclerc". These are names of Formula 1 drivers, therefore this article is likely about Formula 1. Three other words with a higher weight is "race", "somerset", and "essex", which tells us the location of the race.

Tina Hegelund Hyldahl Martin - HSR607
Mikkel Birk Nielsen - FXK415
Kasper Senika Larsen - PHM536
Christian Runge Hansen - QKT122

Topic 3:

In this WordCloud the three words with the highest weights are "world", "players", and "team". These words can be used within a large number of sports. However, by looking further down the list of words, we see words like "football" and "england" are placed as 6th and 7th highest weighing words. This allows us to conclude that this article is about football and probably about the World Cup in football since cup is also one of the most mentioned words.

Topic 4:

In this WordCloud the highest weighing word is "set", which obviously leads one to think that the article is about tennis. There are more words that back this up: "court", "nadal", "serve", "federer", and "williams". Three of these words are the names of prominent tennis players, while the words "court" and "serve" are used often within tennis.

Topic 5:

Lastly, the final WordCloud has names of famous football clubs as some of the highest weighing words; "city", "united", "liverpool", and "arsenal". The highest weighing word; "min" leads us to believe that this article is in fact a live blog post, that discusses some of the top football clubs in England during a game day eg. "at min 5 liverpool are close to take the lead against City".

Tina Hegelund Hyldahl Martin - HSR607
Mikkel Birk Nielsen - FXK415
Kasper Senika Larsen - PHM536
Christian Runge Hansen - QKT122

Bibliography Portfolio 2

Downey, A. (2015). *Think Python. How to think like a computer scientist.* 2. Ed, Version 2.4.0. Needham, Mass.: Green Tea Press. Retrieved August 21, 2019 from http://greenteapress.com/thinkpython2/thinkpython2.pdf

TheGuardian.com (2019). *Open Platform*, set online 9 december, 2019 https://open-platform.theguardian.com/explore/

Ignatow, G. & Mihalcea, R. (2018). *An Introduction to Text Mining : Research Design, Data Collection, and Analysis*. Los Angeles, CA.: Sage.

Lavin, M.J. (2019). Analyzing Documents with TF-IDF. *The Programming Historian* 8

Sweigart, A. (2015). *Automate the boring stuff with Python : practical programming for total beginners* (6th printing.). San Francisco: No Starch Press.