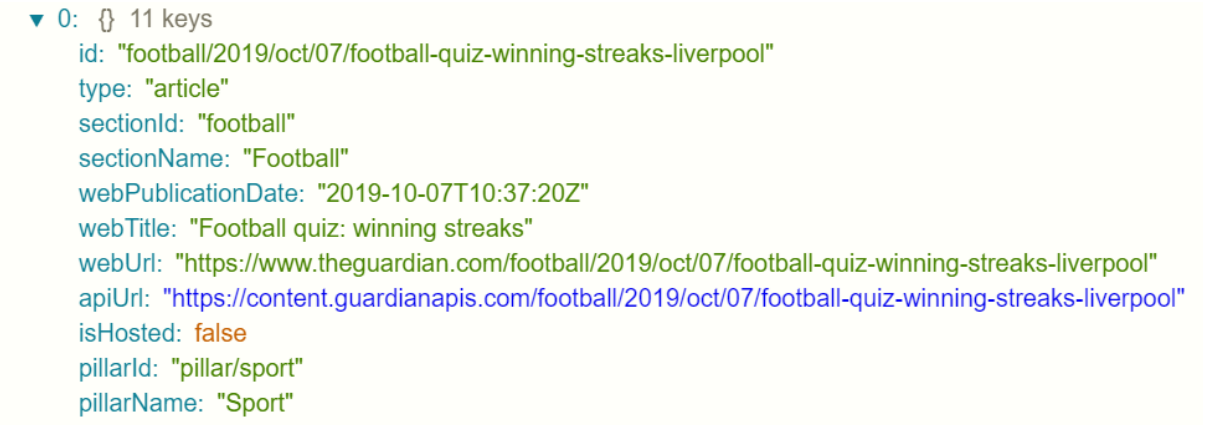
Tina Hegelund Hyldahl Martin - HSR607
Mikkel Birk Nielsen - FXK415
Kasper Senika Larsen - PHM536
Christian Runge - QKT122

Portfolio 2: Key Findings and Reflections

Data selection

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.

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 - QKT122

Hjem - KUnet
Portfolio 2 - Google Docs
theguardian / open platform - e
Football quiz: winning streaks | F
theguardian.com/football/2019/oct/07/football-quiz-winning-streaks-liverpool
Apps
Bookmarks
Google
Syning
Uni
Hår
Mad
Oscar
Other bookmarks
Support The Guardian
Available for everyone, funded by readers
Contribute →
Subscribe →
Search jobs
Sign in
Search
International edition
The Guardian
News
Opinion
Sport
Culture
Lifestyle
More
Football ▶ Live scores Tables Fixtures Results Competitions Clubs
Football
Gregg Bakowski
Mon 7 Oct 2019 11.37 BST
11
Read The Guardian without interruption on all your devices
Subscribe now
most viewed

Football quiz: winning streaks

Liverpool have won 17 games in a row in the Premier League. How well do you remember these victorious runs?

Live General election: Boris Johnson faces Jeremy Corbyn at PMQs amid pre-poll exodus of MPs - live news

Army officer tells impeachment inquiry of gaps in Trump's Ukraine

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.

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.
Some of these words were also among the most used words for the “Football” and “Sport” sections. The other words on the list of most used words for these two sections give a good indication that they are related to football and sports ('one', 'first', 'game', 'back', 'two', 'last', 'time', '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.

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.

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.

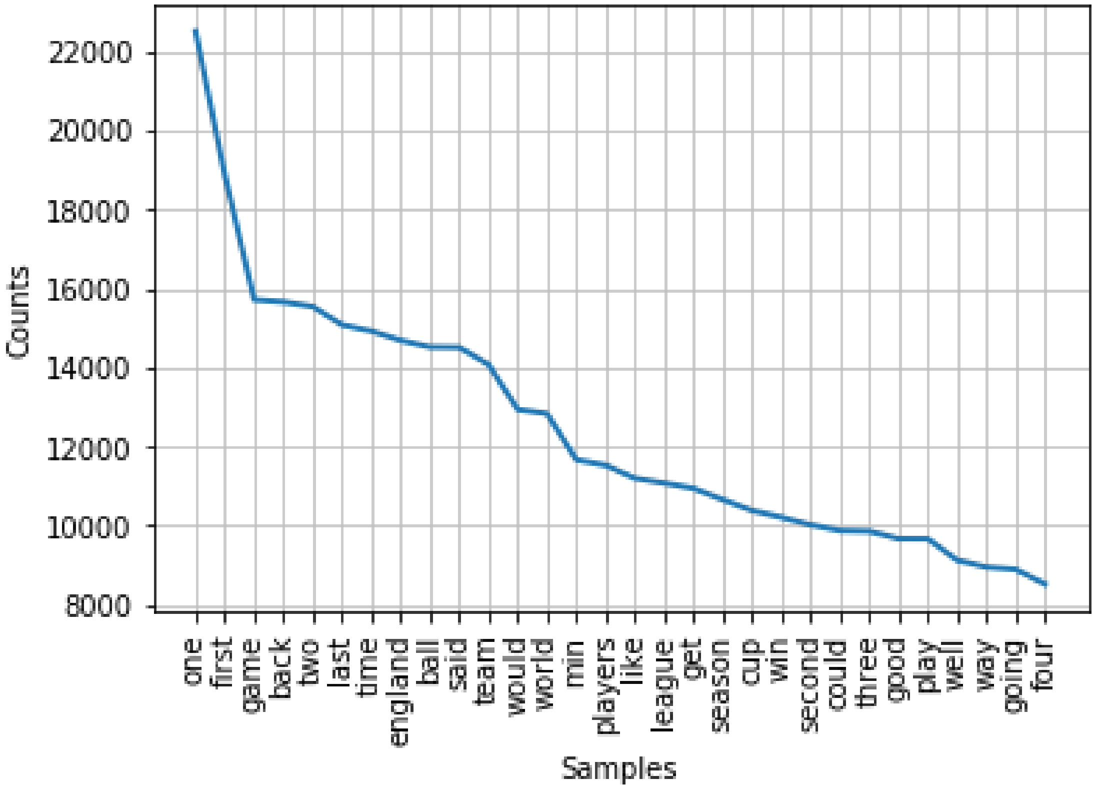
In the following of our research question: With which topics was Tottenham (football team) mentioned? We expect to find the topics: Tottenham, hotspurs, spurs, stadium, hotspur, new, and home. This is because we know that the players on the Tottenham football team are referred to as the “hotspurs” or just the “spurs”.
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.

Model and visualization of topics

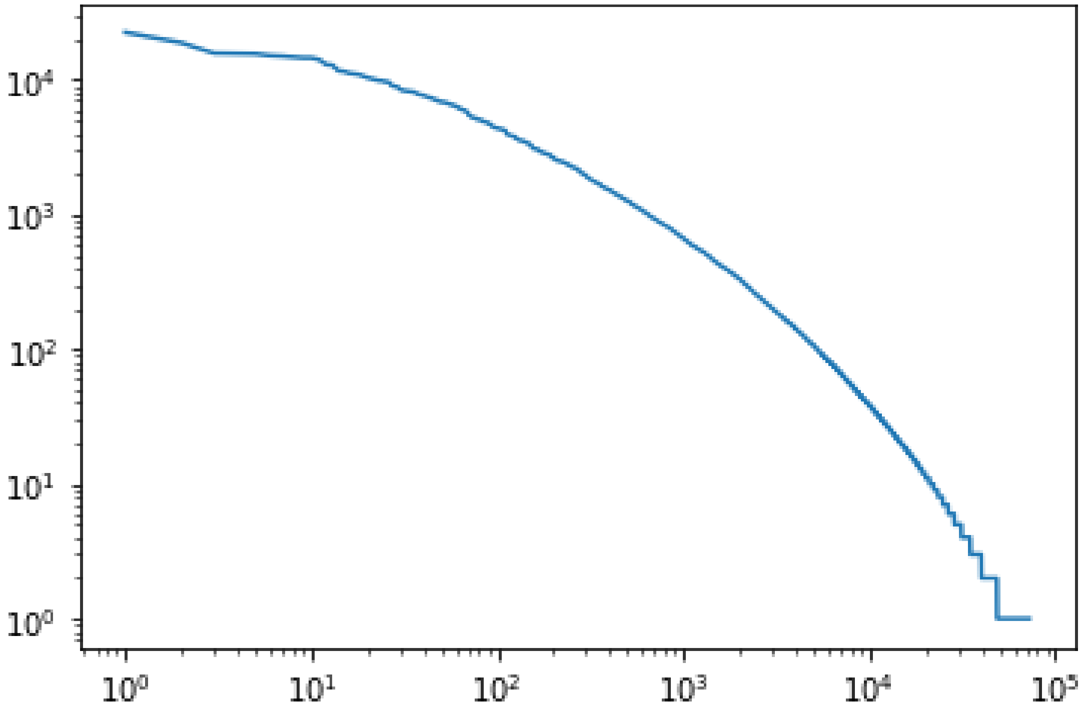
Query terms presented in a tf-idf matrix:

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

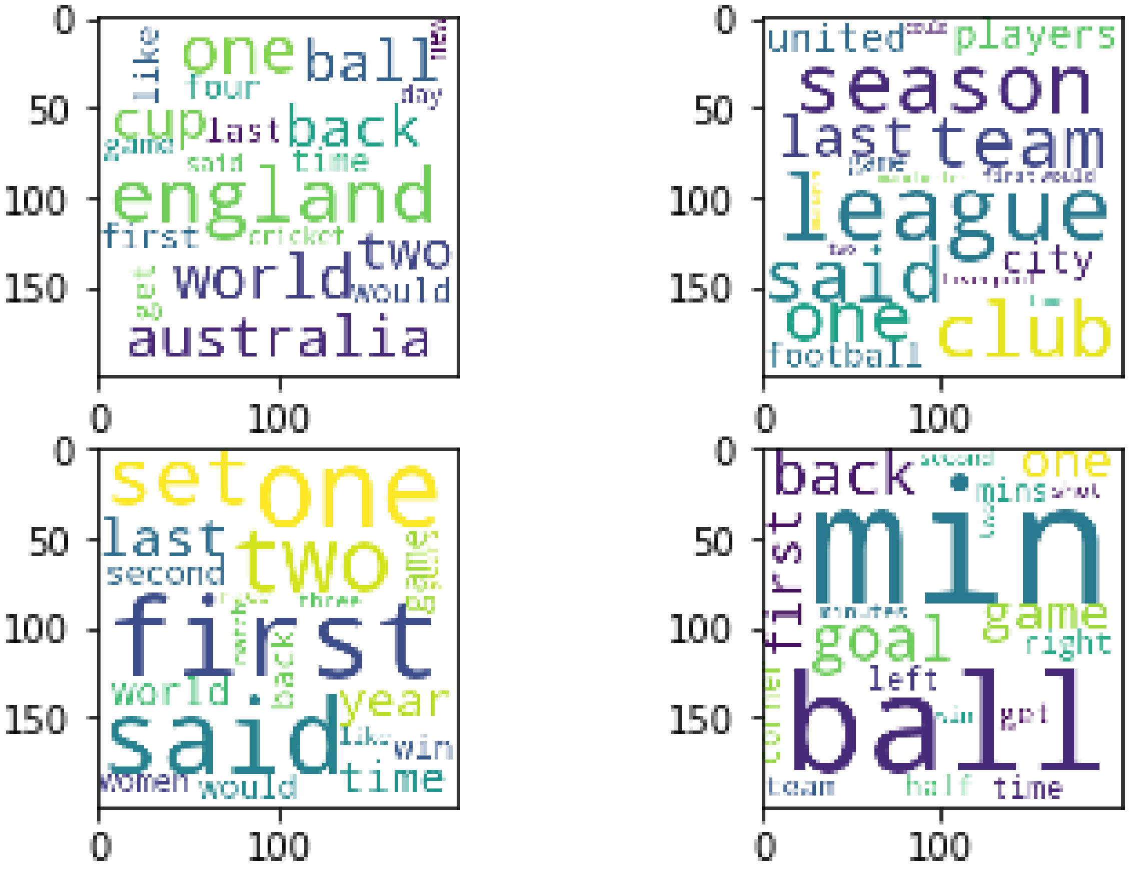
Frequency distribution of the top 30 most used words:



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



WordClouds for the topics with the highest weight from 4 random documents:


one
ball
like
four
day
cup
last
back
game
said
time
england
first
cricket
two
get
world
would
australia
united
players
season
last
team
league
said
city
one
club
football
set
one
last
two
second
game
first
three
back
world
said
year
like
win
women
would
time
back
second
one
mins
min
first
minutes
game
goal
right
left
ball
corner
get
team
half
time