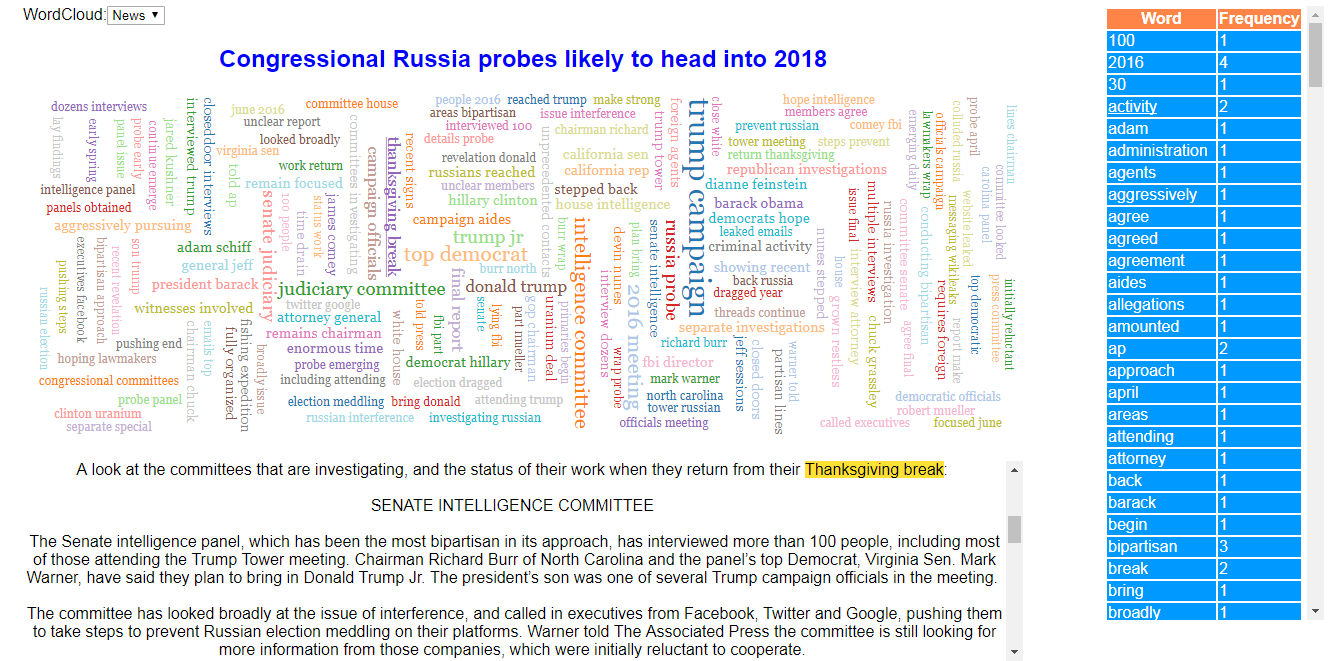
Text Visualization as Word Cloud

Author: Avnish Pandey; Email: [avnishpandey113@live.com](https://d.docs.live.net/399aa50a8ddbb094/Desktop/MyProjects/InformationVisualization/TextVisualization/avnishpandey113@live.com);

## Screenshot:



## Instructions for compiling and running program:

Visit: <http://www2.cs.uregina.ca/~akp817/TextVisualization/html_pages/TextVis.html>

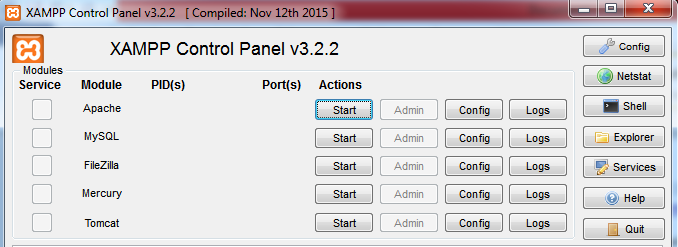
Files: <https://github.com/avnishpandey113/InformationVisualization>/TextVisualization

(OR via Xampp)

Xampp Deployment:

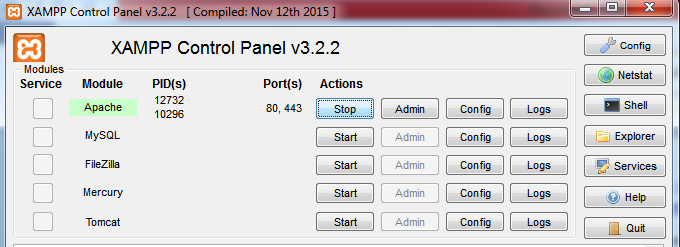
Step 1: Install Xampp from <https://www.apachefriends.org/index.html>.

Step 2: Once installed, open Xampp, below screen will pop-up:



Step 3: Select start in “Actions” column opposite “Apache” module.

Step 4: Apache server will be deployed on your system and the xampp control will look as below:



Step 5: Go to location htdocs where xampp is installed is. It’ll look something like this: “drive:\xampp\htdocs”.

Step 6: Extract all files in this folder from zip file into this folder. Now the folder will look like “drive:\xampp\htdocs\ZipFileName\”

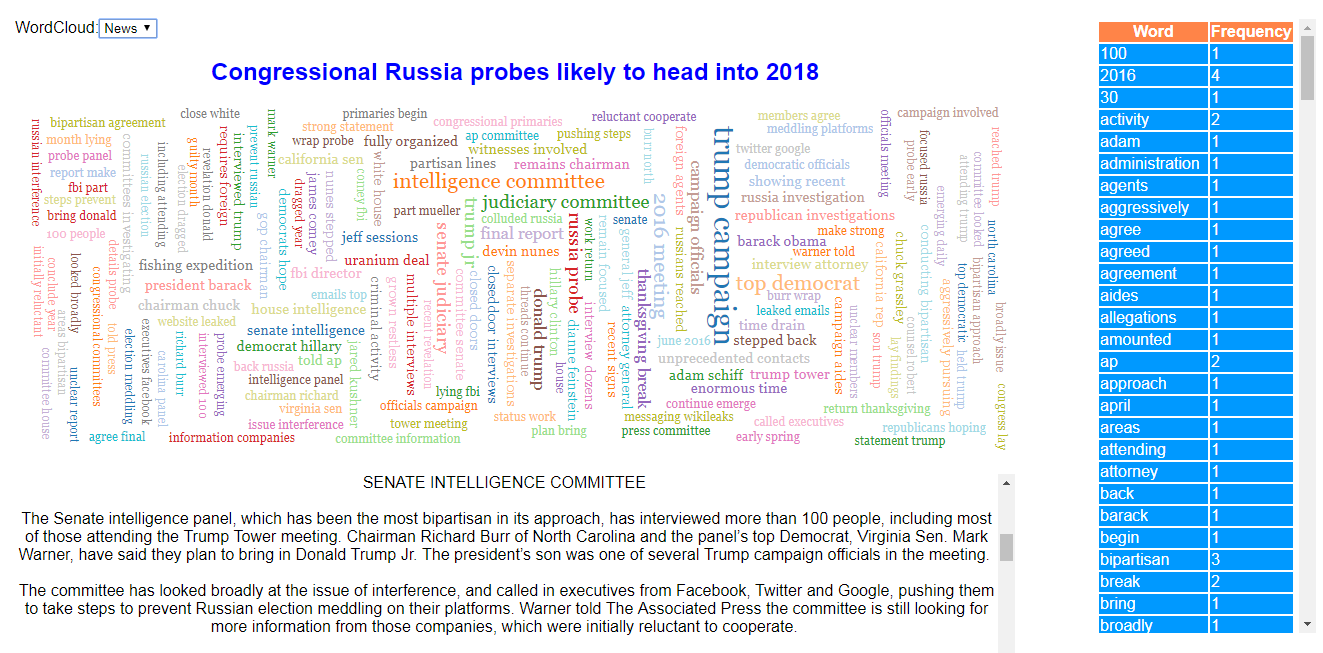
Step 7: Open chrome and type [http://localhost/ZipFile/html\_pages/html\_filename.html](http://localhost/ZipFile/html_pages/html_filename.html%20)  to see the output.

## Discussion:

**Explanation of Visualization:**

1. The visualization represents the co-occurrence of the pair of terms in the form of word cloud for a news and a short story.
2. The user can select what to represent as the word cloud using the drop-down menu provided. It has 2 options, one for news and the other for story.
3. The news taken is between 500-1000 words and the story is between 2000-4000 words.
4. During representation of the co-occurrence of terms, the pair is taken such that there is no occurrence of stop words in any of them.
5. Random pair of words are selected from the array and shown on the world cloud for easy readability. But the ones with the higher frequency will always be visible.
6. Frequency of the pair of words is represented by the size visual variable. Higher the frequency, bigger the size of the pair.
7. The color visual variable is used to represent each pair of term occurred in the document so that user can easily distinguish between different pair of terms.
8. The visualization represents the frequency of each individual term in the document as well in the form of table. One column names the word and the other its frequency.
9. The user can find the occurrence of each individual term in the document just by clicking on the word in the provided table.
10. The word clicked will be underlined and the document will auto-scroll to the point at the first occurrence of the word inside the document. The words will be separately highlighted with green color within the document as well for easy identification. If the word has occurred multiple times, it’ll highlighted at multiple places in document.
11. The user can find the co-occurrence of the pair of term in the document just by clicking on the word in the word cloud.
12. The clicked pair will be underlined and the document will auto-scroll to the point at the first occurrence of the word inside the document. The pair will be separately highlighted with yellow color within the document as well for easy identification. If the pair has occurred multiple times, it’ll highlighted at multiple places in document.

**Visualization for shorter documents:**

****

The above screenshot represents the visualization for short documents, in this case it’s of news. We can conclude the following points from it:

* The co-occurrence of pair of terms is clearly distinguishable.
* The total number of words represented are less.
* The frequency of the pair of terms is clearly distinguishable by the size of the pair.

**Visualization for longer documents:**

****

The above screenshot represents the visualization for long documents, in this case it’s of story. We can conclude the following points from it:

* The co-occurrence of pair of terms is not so much clearly distinguishable.
* The total number of words represented are far more than the shorter document.
* The frequency of the pair of terms is not so much clearly distinguishable by the size of the pair.

**Benefits:**

1. Word clouds are graphical representations of word frequency.
2. They give more importance to the words that appear more frequently in the document.
3. The bigger the size of word/word-pair, the more common it is in the document.
4. This type of visualization can assist evaluators with exploratory textual analysis by identifying words or pair of terms that occur frequently in a set of interviews, documents, or other text.
5. It can also be used for communicating the most salient points or themes in the reporting stage.
6. One should remove the stop or common words before using this approach for text visualization.

**Drawbacks:**

1. The word cloud emphasizes frequency of words, not necessarily their importance. Word clouds will not accurately reflect the content of text if slightly different words are used for the same idea (for example, ‘large’, ‘huge’, ‘giant’, ‘enormous’, and ‘big’).
2. They also do not provide context of the document, so the meaning of individual words or pair of words may be lost.
3. With the increase in the size of the document, the visualization of the words becomes more congested. So it becomes hard to visually distinguish among the words.
4. One cannot make sense of a complex topic like research paper or war just by looking only at the words used to describe the event.

**Features User Can Find Among Documents:**

1. Words that have occurred more frequently in the document.
2. User will get an approximation of what the document is saying by identifying the pair of terms or words in the visualization.

**Aspects of Documents Difficult to Observe:**

1. If the same idea is conveyed using different words in the document, the user will not be able perceive that particular idea.
2. The user will not be able to make complete sense of the document just by looking at its word cloud.

## References for code:

* <https://bl.ocks.org/blockspring/847a40e23f68d6d7e8b5>
* <https://stackoverflow.com>