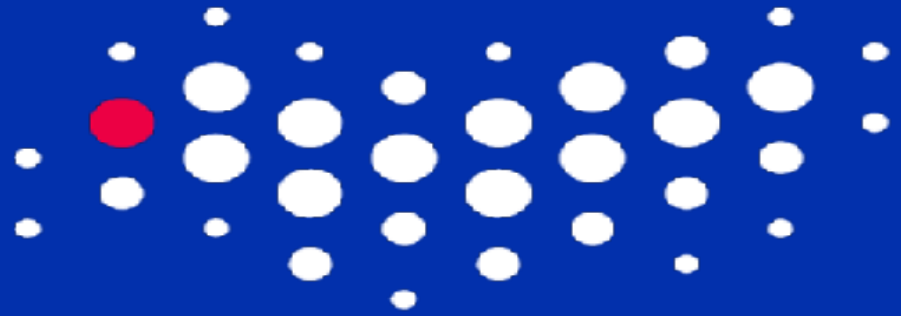


ITMO UNIVERSITY

Visualisation of conversation as a semantic network.

Nicholas Amoako

Introduction

This research is about designing graphical representation for dialogue between a simulated interactive software called Pythia which mimic the behaviour of the ancient Greek high priestess (called Pythia) in the temple of Apollo (an ancient Greek god). Pythia was highly regarded for it was believed that she channeled prophecies directly from Apollo to tell people about their destiny in this world. Pythia was not just an oracle but was regarded as both a political and spiritual figure in Greek history. It was believed that prophecies from Delphi helped shape modern civilisation and defined the course of history. She was also known as the Oracle of Delphi and was consulted in all major decisions before action was taken. Many rich and influential people pay a fortune to consult the oracle and it was believed that even the gods also sought advice from the oracle.

Although few people remember her and still believes her legendary powers, her wisdom continues to inspire us and teaches us that through knowing yourself, you’re able to penetrate the mysteries of the past and the future.

Aims and Objectives

We will aim at designing an effective symbiosis between human perceptual and cognitive skills and the computer ability to mine and summarise text to design our interface by following these stated objectives:

- Create a visual representation that highlight relevant data and sentiment in conversations to help people make sense out of the social dialogue archive.
- To develop a system that is able analyze and visualise any text conversation from any social media platform(live feed and Archives) to derive sensitive information and recognise social patterns that that can serve any purpose in terms of security, market prediction, health and others.
- To adopt a simplistic strategy to base the interface component in common metaphors so that the interface can be used by a large user population.
- Although we are basing our case study on the Pythia dataset, we will expand our algorithm information extraction, analysis and visualization to be applicable to other works on different conversational modalities such as emails, blogs, etc

Related Works.

Our approach is not to just create charts and maps from the text as seen in many documents and tools but rather preprocess the data to draw a semantic graph which gives a different direction and insight into the data. The sources of literature for this project was selected from science direct database. This database is well known for high end repository of peer reviewed literature for scientific and medical research.

We utilised keywords such as: semantic network, visualisation, text data graphs, social media dialogues and retrieved over 3,000 peer reviewed papers and journals in the past five years.

About 30 papers were reviewed in the following subject areas: Visual interface for text analysis, Text analysis as a semantic Network, Implicated Meanings in text conversations, Making sense of complex data through visualization, Text analysis using Artificial Neural Networks, Automatic Platform for text and sentiment Analysis, Pattern of conversation in social media

Technologies Used

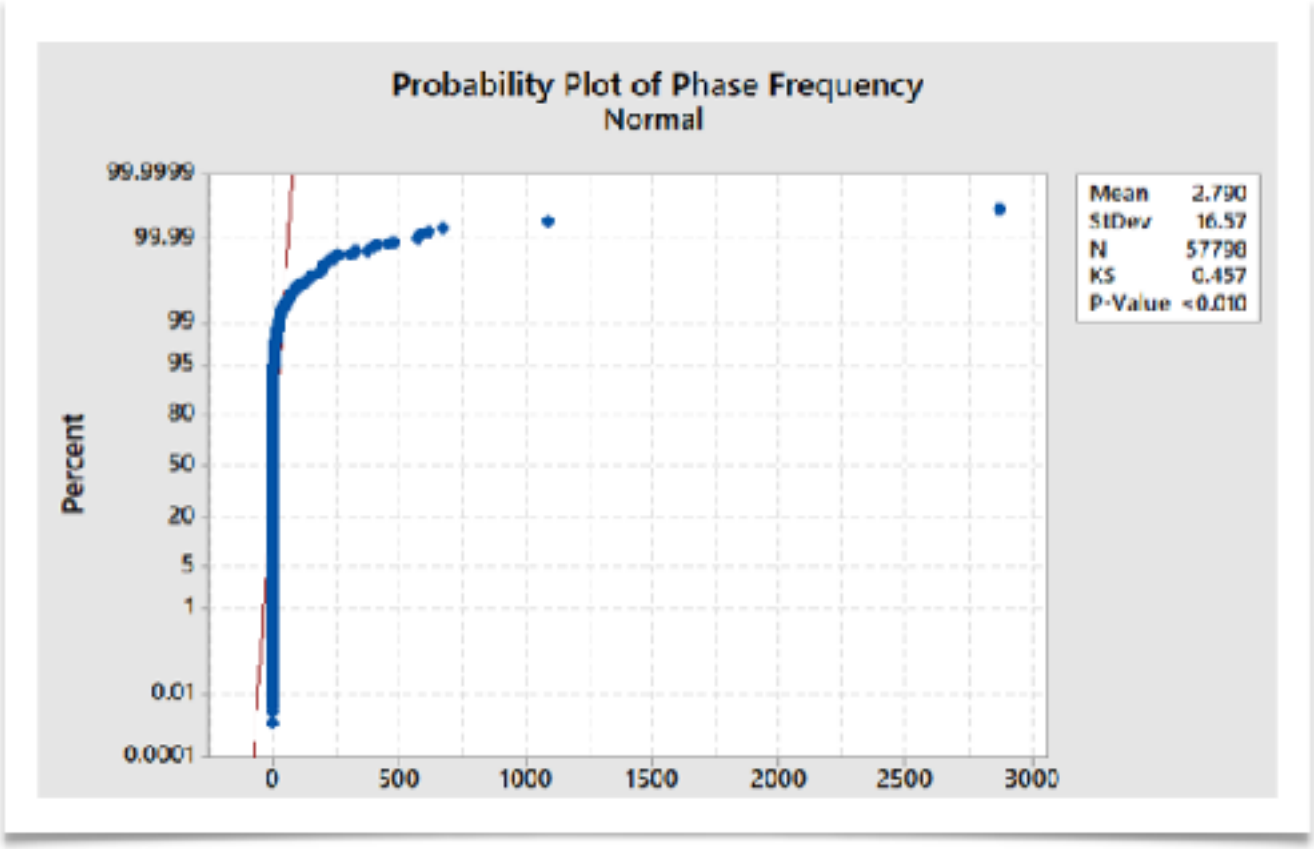
- Nodejs
- Express framework
- Anaconda Notebook(Python)
- D3 js
- HTML
- Javascript

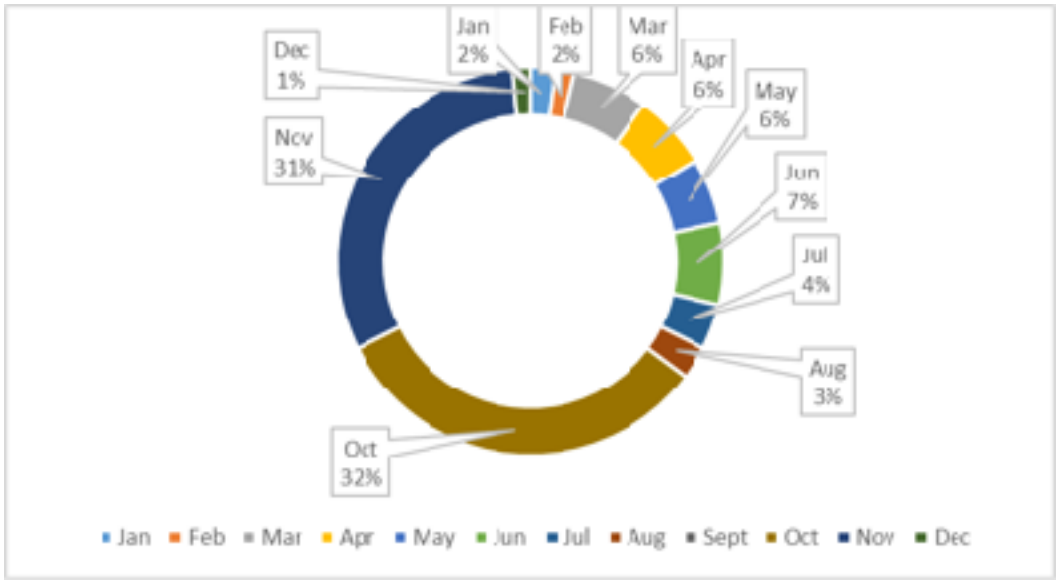
Dataset Description:

- The data was gathered from an interactive installation called Pythia through google speech API. [http://www.sonicartist.me/wp/pythia/] It is an interactive software that analyse audio speech into a prophetic message. This was designed to mimic the ancient Greece Goddess of Apollo. By the use of microphone, people's voice was recorded as they talk to the "oracle". The data collected covers a specific time frame from Jan 2015 to December 2015.
- The dataset was a csv file with 7 columns (Number of Records (bin); Number of Records; Client; Id; Lang; Text; Time) and 161533 rows. Our target aim allows us to select only the "text" and "Time" fields from the dataset.
- Jupyter Notebook was used in conjunction with pandas' library in python to remove invalid fields and erroneous records from the dataset. After the cleaning of the data, the data was group according to month thus: from January to December and an additional field was created called "count" to take into accounts how each phrase appears in the dataset per month.
- The year and days parts of the date was striped since we are only interested in grouping the data according to phrase count and months and also the all fall under one year (2015).

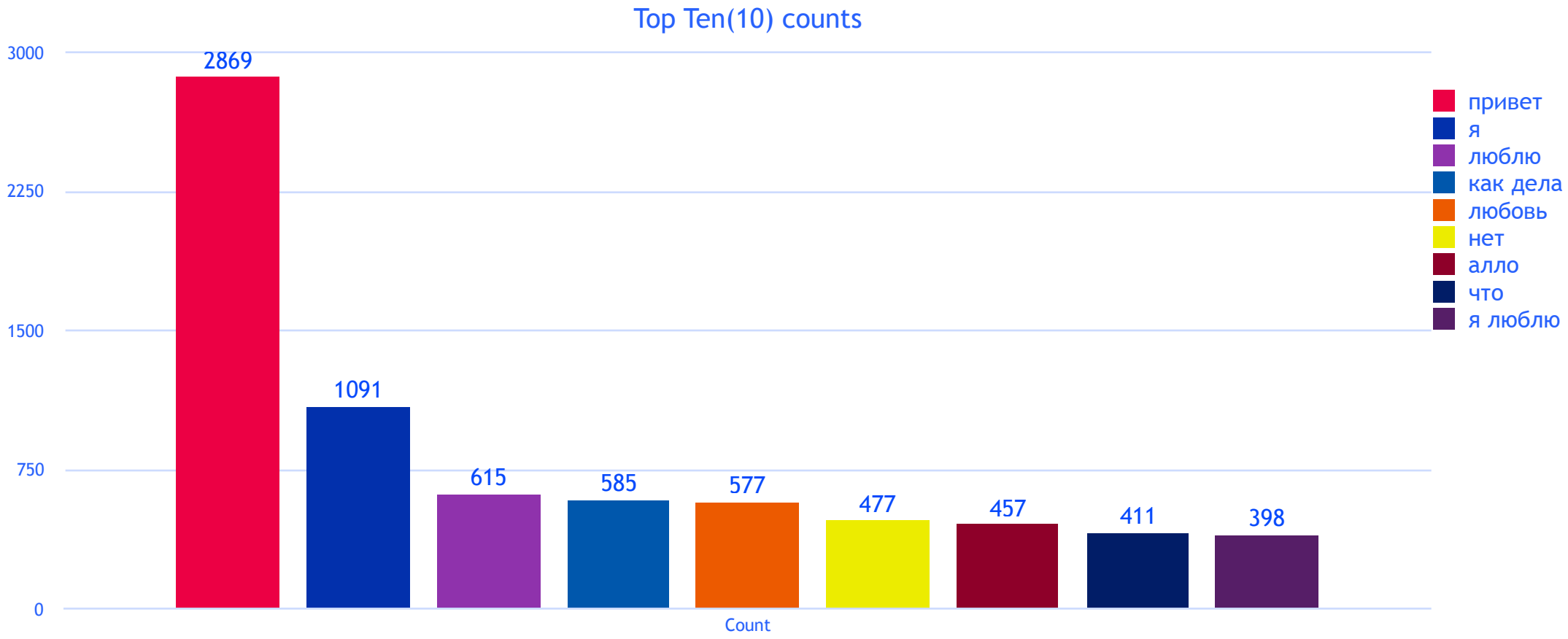
Data Analysis

- We analyzed the frequency variations in the phrases and attained a mean value 2.7 and standard deviation of 16.6. the p value was less than 0.01 which signifies that the data is not normally distributed.
- The figure demonstrates that about 99% of the have appearance between 1 and 20 leaving only 1% of the phrase having greater representation. An outlier was discovered with one phrase appearing almost 3000 times in the dataset.
- Therefore, we can confidently say that there is a high amount of unique phrases in the dataset.




Jan 2%
Feb 2%
Mar 6%
Apr 6%
May 6%
Jun 7%
Jul 4%
Aug 3%
Oct 32%
Nov 31%
Dec 1%
Jan Feb Mar Apr May Jun Jul Aug Sept Oct Nov Dec

Top 10 most frequent phrases.



Data Visualization

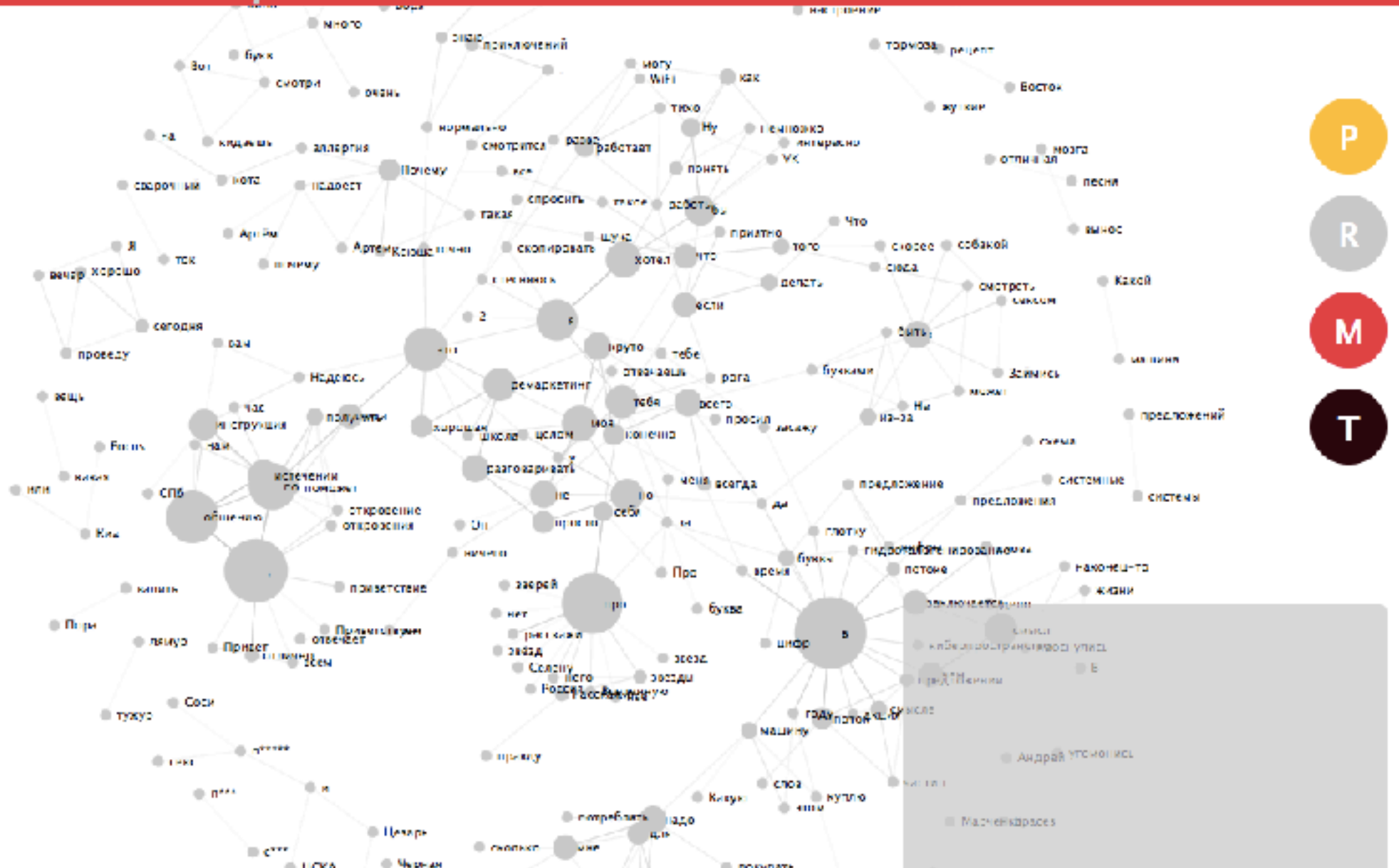
- D3.js Library was the chosen data visualisation tool for our project.
- There is summary view which consist of the complete semantic network of disjointed words connected with arrows showing the direction and order of precedence.
- create an ontological view that provides a systematic way for the users to explore the relevant concepts in the conversation and their relations.
- Svg circles are drawn to indicate each node of word in the phrase.
- An interconnecting line that establishes the relationships between nodes
- A third view called transcript view will display original phrase in the network with specific color (e.g.: blue) when clicked and other evolving phrases that has been generated due to their interconnectivity in the network (e.g.: yellow). If entities are selected, the main node and its tributaries will be highlighted with their corresponding color described earlier in this paragraph.

ITMO UNIVERSITY

Pythia Data Visualization

search

name	count
2	1
работает	7
тихо	3
если	9
работы	3
WiFi	2
ляция	3
в	20
мысле	6
поток	8
частиц	4
предложений	5
Не	2
быть	4
из-за	6
может	4
смотреть	4
слюда	4
того	6
что	9
буквы	6
за	4
буква	1



ITMO UNIVERSITY

Pythia Data Visualization

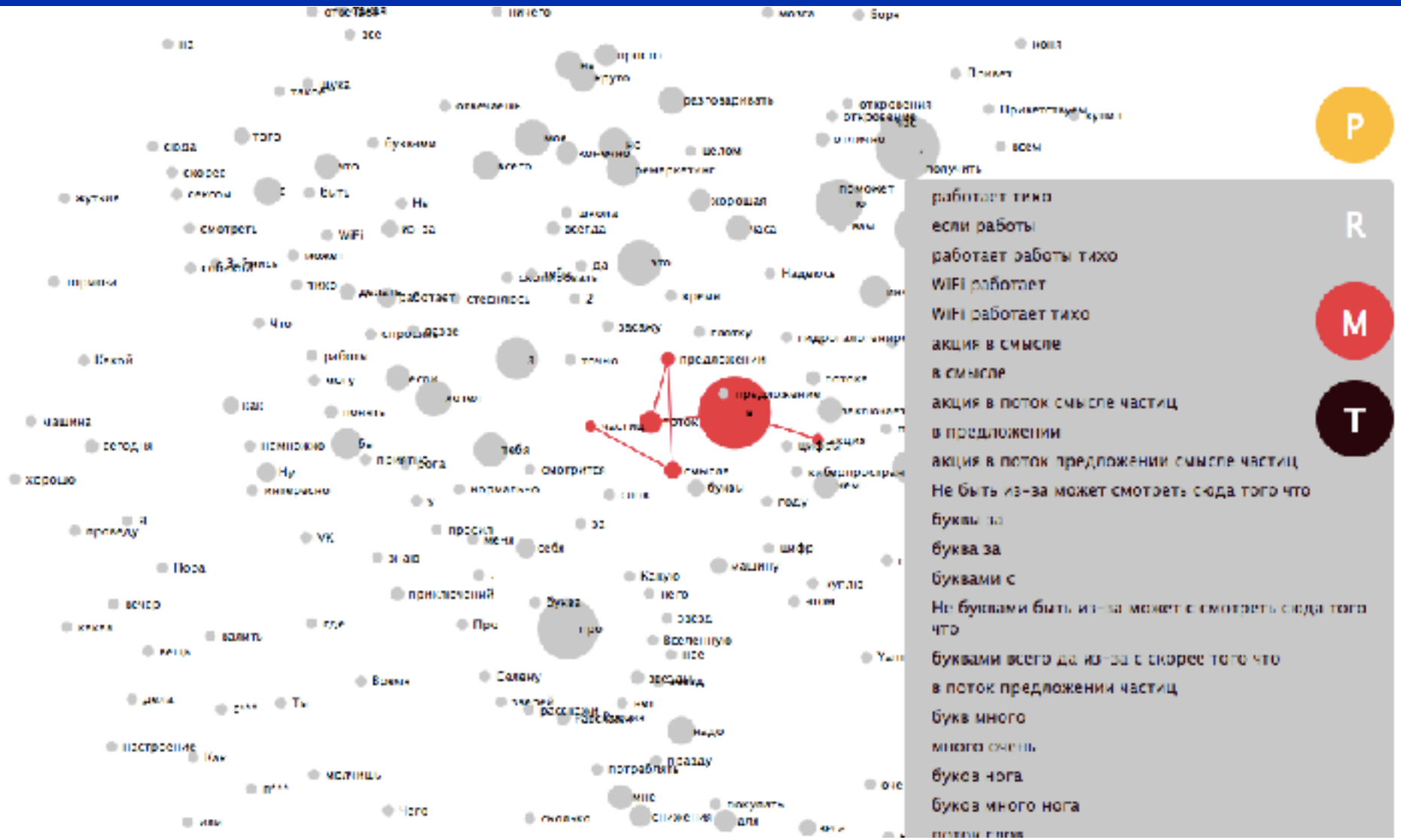
search

name	count
2	1
работает	7
тихо	3
если	9
работы	3
WiFi	2
акция	3
в	26
смысле	6
поток	8
частиц	4
предложении	5
Не	2
быть	4
из-за	6
может	4
смотреть	4
сюда	4
того	6
что	9
буквы	6
за	4
букв	1

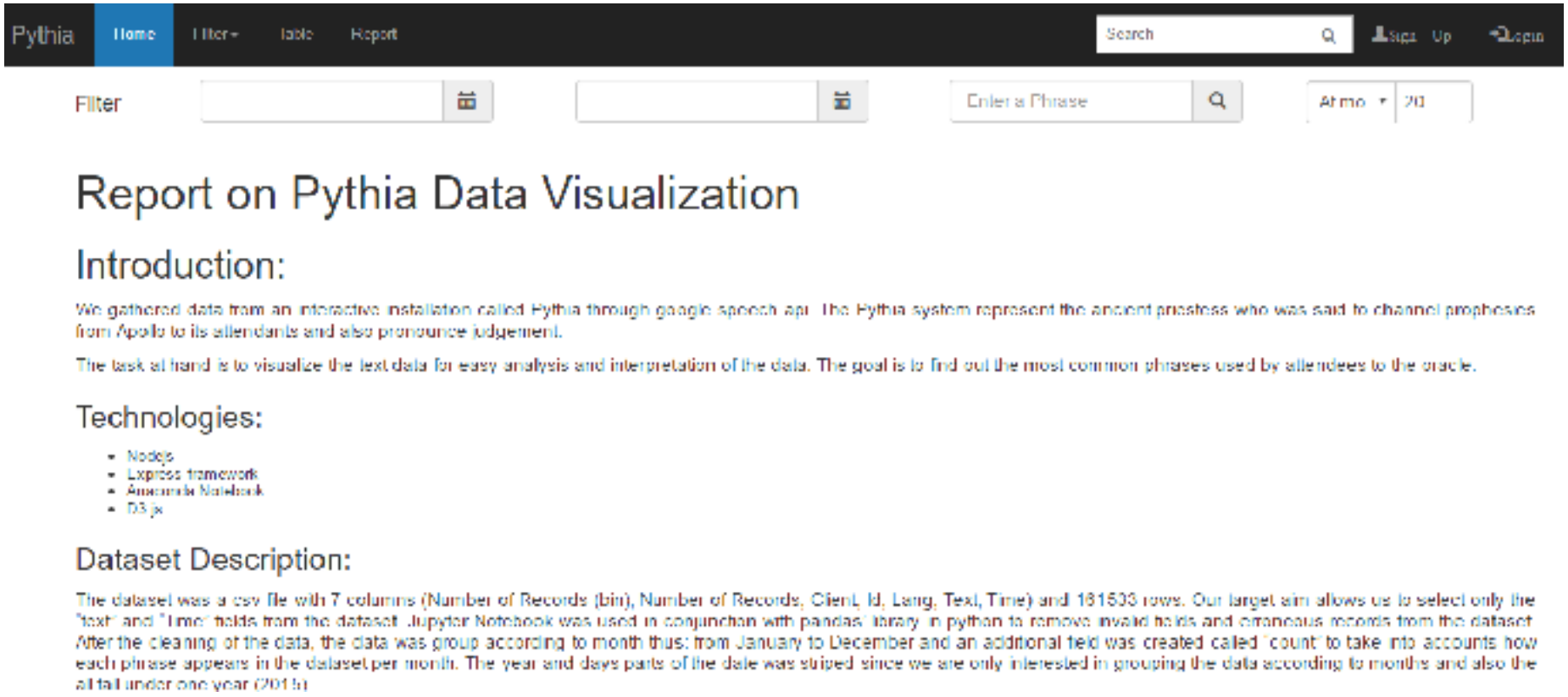
работает тихо
работает работы тихо
WiFi работает
WiFi работает тихо
Ну как могу работает скопировать это я

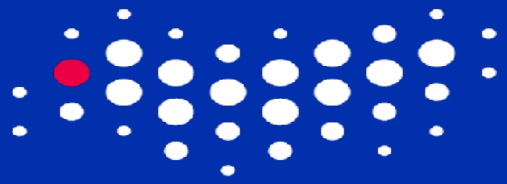
search

name	count
2	1
работает	7
тихо	3
если	9
работы	3
WiFi	2
акция	3
в	26
смысле	6
поток	8
частиц	4
предложении	5
Не	2
быть	4
из-за	6
может	4
смотреть	4
сюда	4
того	6
что	9
буквы	6
за	4



Report Screen





Further Studies

- To find the trend in the data in terms of connecting the phrases based on their time lags.
- This is because it can be assumed that one phrase can be divided into different sentences due to the delay in the speech.

THE END