

Story (20)

– A ne pas négliger, 20 points

Motivation:

- Why graphs ? Nice representation of some data that cannot be represented otherwise, ...
- Why Wikipedia ? nice graph structure,
- Why Graph ML ? : current trend, interdisciplinary, take advantage of graphs
- Why recommender system ?

Sell our product

Relevance of chosen data and tools:

- Why wikipedia and Seealsology ? : not to get a fully connected graph, ...
- Why Glove, fast text >> **TBD**, NLP etc...
- Why Deepwalk >> Graph ml trends etc...

Une demi-page de rapport:

- (10/12) @Ayyoub Parler de ce qui est déjà décidé
- To be finished for the report

Acquisition (10)

(10/12) @Vincent

Graph construction with <https://www.analyticsvidhya.com/blog/2019/11/graph-feature-extraction-deepwalk/> : Make in modular to choose bigger and bigger graphs

Collect data through Seealsology, clean it : make it complex in appearance (bonus points ^^)

Combining datasets : take keywords from github dataset instead of title page ?

Exploration (20)

(10/12) Frédéric & Ayyoub

Create a script that given a graph gives:

- Number of connected components
- Diameters
- Clusters Degree Distribution
- Clustering coefficient
- Centrality
- Distribution smoothness of the attributes
- NLP simple analysis of the attributes (word count etc...)

Identify type of graph – get ready for (11/12)

A visualization of the graph TBD (11/12) <https://plot.ly/python/v3/3d-network-graph/>
Run code (12/12)
A reflection on the insights **TBD**

Exploitation (20)

- Spectral Clustering (TBD on which adjacency matrix) @Karthigan (10/12)
- PCA, t-SNE on the embeddings @Karthigan (10/12)
- DeepWalk @Vincent (10/12)
- Critical Evaluation of the results (17/12)
 - Remove a node and find it, use query expansion through another Word2Vec model **TBD**
 - Subjective : Google Search
 - Limitations **TBD**

Communication (20)

Report (From 16/12)
Interactive Visualization (From 11/12) – I will start thinking about it @Karthigan

For 12/12:

Story, Acquisition, Exploration, Exploitation (without results) finished

From 12/12 to 17/12 Results

From 12/12 to 19/12, Interactive Visualization

From 16/12 to 20/12, Report

20/12 Cleaning Repository

20/12 Report, Interactive Visualization, Repository Finished

Every results should be in one notebook, the functions code should in python script properly commented with docstring.

The final interactive visualization that is the final product is a new notebook