Alina Sulkovsky 316015247

Guy Shafir 315941799

Eylam Kaddan 206516957

Ron Hugi 313470189

Liam Ashkenazi 211890314

TOPIC MODELING WITH NLP

Data Science Two-semester Project

LDA Model

BACKGROUND

Many organisations have found topic modelling as a useful way to explore large text collections. Unfortunately, running customised models usually requires a dedicated team of data scientists, and it can take some time before there'll be useful results.

MAIN GOAL

Expectations and outcomes

The goal of this project is to make running topic models easy for anyone with a modern web browser, with all the preparational work like gathering data and training models, done in advance.

That way the team can focus on analysing results and adapting the organisation working flow accordingly.

STAKEHOLDERS

Due to the generic nature of the project, it may suit the needs of professionals from a great variety of fields. The source of the data is free public information on the internet, therefore all firms related to the public sector are a potential client. For example: press, law firms, municipalities and more.

PROJECT STRUCTURE

PYTHON BASED APPLICATION

ELASTIC SEARCH

For managing html data scrapped from massive-public information on the internet

PYTHON CODE

Using python we are cleaning and normalizing the data from elastich search, saving it as mongoDB documents, applying LDA model for the topic modeling

MONGO DB

Managing data collections of the source data and analyzed data

Main Steps

Web scraping

Retrieve HTML source data from a list of predefined links, including public mass information

Data Cleaning and Normalization

Using custom cleaning function and Hebrew-NLP library we are preparing the data for the model. Saving the source and proccessed data as saparated collections in MongoDB

Finding Bigram vs. Trigram For Sanity Check

Checking that the data we colected makes sense and that we can proceed to the topic modeling process

Applying LDA Model

Using tomotopy LDA Model we get the Topics existing in our data, saving it as a collection in MongoDB for later use in a node graph

Conclusions

Team Work:

Goal setting, splitting tasks, task synchronization, shared learning

Deep Knowledge:

During shared work on the project we feel that we had the right to work with professional tools such as:

Python

Elastic Search

MongoDB

LDA Model

etc...

Technology Stack

HebrewNLP by INFONETO

mongoDB

python™

elasticsearch

THANK YOU