Document summarization for a knowledge product company

Client Overview

- A market research knowledge product company based out of the USA

Business Challenge

- 120mn+ documents scraped and stored in a Postgres database. The documents were of different types - research papers, market research documents, blogs etc.,
- The vision for the product was to present only relevant documents to the users according to their interest
- The target audience was the key executives who had time enough to read only summaries

Implementation Approach

- Reading the whole documents would have taken a lot of executive's time
- Reading one document might not have given them the complete picture of the market or the search topic as one document might focus only on one subject
- There was a need for a method which could extract brief summaries, key topics, entities and quality of publishers etc.,
- Topic Modeling on retrieved documents from Elasticsearch using Latent Dirichlet Allocation algorithm. Topic models presented as network charts using D3.js
- Named Entity Recognition using SNER and SpaCy
- Publisher-publisher similarity using word2vec, TF-IDF & Cosine Similarity

Technology

- LDA, Word2Vec, Cosine Similarity (Hetzner machines, Flask, Docker, SpaCy)

Business Impact

- Understand high level topics in minimal time
- Publisher recommendations and quality check
- Develop high level understanding of several topics and find trending topics