Key Concept Analytics Problem

Introduction

One way of understanding and speeding up the consumption of unstructured text by end-users is to automatically extract so-called “key concepts” and order them in a meaningful way. Performing analytics over such extractions becomes a key task to facilitate knowledge acquisition and curation.

Your Task

You will operate on an in-house collection of patent documents. It is your task to parse these patents and to semantically enrich them using key concept/phrase extraction so that users can later quickly review documents by looking only at them instead of reading the text. To support that, you should store the key phrases and document identifiers in a data structure or database that supports fast retrieval for analytics purposes. As an additional use-case, please generate a meaningful ordering of the top-30 key phrases also for every document that could help to understand the main theme.

!!!The description is fuzzy on purpose, i.e., we didn’t specify any details for extracting keyphrases or for the data storage paradigm. Please make pragmatic choices for these underspecified parts.!!!

Of course, we don’t expect you to create a perfect system, but we are very curious about your ideas for further improvement – even if you decide to not implement them. **We really ask you to include appropriate comments in your source code that help us understand your design decisions while we review your code.**

Some clarifications:

- The input is a set of archives containing XML documents. The outer zip exists only for technical reasons.
- Apply the key phrase/concept extraction on the abstract of each patent
- You can use the filename as the document identifier
- Your solution should be reusable (i.e., by being wrapped in a Docker container)
- We have only provided some sample data – will your solution be able to also process hundreds of millions of documents?
- You can collect all ideas for further improvements and/or all assumptions in a README

You can download the data here: https://databricksexternal.blob.core.windows.net/hiring/patents.zip?sp=r&st=2021-10-07T23:09:03Z&se=2021-10-31T08:09:03Z&spr=https&sv=2020-08-04&sr=b&sig=uR36HP3kCEDY9aPc0mvZFzLnblodA9adxQRTYTc6O6M%3D

If anything is unclear, please reach out immediately. We will act like your customers. Feel free to clarify the needs via emails or by setting up short calls with us. As data scientists, we must often engage with our customers to understand their needs. Always the specification is wage.

Deliverables

Please provide us with:

a. your code as GitHub repo (otherwise as local git repo via email).

b. a readme explaining how to easily deploy/run your UI.

c. a brief explanation of your solution.