Study of classification and summarization techniques applied to the BigPatent dataset

Marco Braga, 829295
Andrea Maver, 828725

Text Mining and Search Project

January 2022

Overview

1. Introduction
2. Pre-processing
3. Texts Representation
4. Classification
5. Summarization
6. Conclusions

Introduction

Goal

Our goal is to develop a useful process to be adopted when a new patent is published: initially, analyze its contents to classify it and put it in the right category, then create a short summary to understand its subject without reading the whole document.

We tried different algorithms and techniques to find the most suited for this tasks.

Data

The chosen data is from the BigPatent dataset, which contains over 1.3 million of U.S. patents. Because of the high computational costs, only a subset made of around 45 thousand documents has been considered.
Here is an example of the dataset:

abstract	description	topic
The invention discloses a device for determini...	The border processing regarding a tissue or re...	g
A dew preventing mechanism prevents the format...	Fig1 a is a sectional view of an air condition...	f

Categories

The patents can be assigned to one of the following nine categories:

- A: Human necessities
- B: Performing operations, Transporting
- C: Chemistry, Metallurgy
- D: Textiles, Paper
- E: Fixed constructions
- F: Mechanical engineering, Lightning, Heating, Weapons, Blasting
- G: Physics
- H: Electricity
- Y: General tagging of new or cross-sectional technology

Pre-processing

Pre-processing 1/2

The first step of the analysis is cleaning the texts, applying the following operations:

1. Check if there are any null texts or abstracts
2. Case Folding
3. Remove URL
4. Remove punctuation (except points and commas)
5. Remove numbers
6. Remove stop-words and terms with less than three characters

Tokenization and Stemming/Lemmatization

We decided to apply different strategies in order to evaluate the possible different results:

1. Tokenization with white space and Porter Stemming
2. Tokenization and Lemmatization through *SpaCy* and Part Of Speech tagging

Pre-processing 2/2

The summarization part required a slightly different pre-processing.

First of all, we only take 2500 patents belonging to the category B, because of the high computational costs of the task.

The documents are divided into sentences through a sentence tokenizer, then cleaned as before, removing the unwanted parts.

Texts Representation

Texts Representation

Two strategies for text representation:

1. Tf-Idf with different number of features and different type of n-grams (1 or 1 and 2-grams)
2. Doc2Vec with vectors' length 350 and trained for 30 epochs

Classification

Classification 1/2

Different methods for classification:

1. Decision Tree
2. Random Forest
3. KNN
4. Linear SVM with SVD

To evaluate their performances, we utilize vanilla accuracy, a variation of standard accuracy.

Classification 2/2 - Results

The following results are obtained with 15000 features for Tf-Idf representation:

	(1,1)	(1,2)
DT	0.488	0.488
RF	0.632	0.635
SVM	0.262	0.328
KNN	0.615	0.626

Table: Accuracy with stemming

	(1,1)	(1,2)
DT	0.497	0.494
RF	0.633	0.637
SVM	0.313	0.302
KNN	0.615	0.627

Table: Accuracy with lemmatization

Summarization

Summarization 1/2

We tried two techniques of summarization: extractive and abstractive

- Extractive: uses a graph representation to represent the document's sentences. Each sentence is assigned a PageRank score and the summary is built with the top four
- Abstractive: employs Pegasus, a natural language processing model from Google to create new summaries

The abstracts are evaluated using the Rouge-1 and Rouge-l metric

Summarization 2/2 - Results

	Rouge-1	Rouge-l
Extractive	0.241	0.198
Abstractive	0.340	0.294

	Summary
Extractive	another embodiment for the provision of the citric acid solution 30 in the first compartment 14 is illustrated in fig4 and is particularly useful when the fluid in the container means is a carbonated beverage as the generation of the carbon dioxide gas continues will provide space for further expansion of the expandable pouch means 2 . the container 40 is illustrated as being a tube but it is to be understood that it can be of any desired geometrical configuration .
Abstractive	An expandable pouch means comprising two relatively flat sheets of a flexible plastic material in superposed relationship and formed into a first compartment and a plurality of other compartments by a plurality of lengthwise extending strips which join together opposed portions of the flat sheets using a semipermanent pressure- rupturable sealing means.

Conclusions

Conclusions

Some of the classification models gave some promising results, that could be improved with higher computational resources and more time.

Abstractive summarization seems more appropriate for this type of dataset, using Pegasus or maybe even a new custom model exploiting all the data available.

Thanks!