Bases de dades avançades
Pràctica 3

Pràctica Curs 2017/18
12 desembre 2018

1 Objectius

Aprendre a treballar amb Spark i tecnologies de visualització avançades

2 Preliminars

Per rememorar l'aniversari de *El Quixot de la Manxa*, volem estudiar la semblança semàntica entre els seus paràgrafs. Per tal de realitzar aquesta tasca, utilitzarem *Spark* 2.1 amb la llibreria ML.

La idea general es convertir en elements matemàticament comparables les paraules i paragrafs per poder assimilar-los a nivell compartiu.

Una de les tècniques per realitzar això és la vectorització de paraules i textos que proveeix spark amb algoritmes com Word2Vec.

https://www.gutenberg.org/cache/epub/2000/pg2000.txt

2 Guió

Primera part de la pràctica

La primera part de la pràctica consistirà en processar el text del quixot amb l'objectiu d'obtenir el model vectoritzat de les paraules i el model resultant de comparació de paràgrafs.

Caldrà fer un preprocessat de les dades tenint en compte:

- Majúscules i minúscules
- Tokenització
- Accents i símbols

Una vegada tenim el *DataFrame* preparat, caldrà entrenar el model de l'algorisme *Word 2 Vector* i tornar-lo a aplicar al llibre per obtenir el vector dels paràgrafs.

Cal doncs que realitzeu:

- Clúster de test a DataBricks
- Importació text El Quixot
- Desenvolupament mitjançant el notebook del procés de neteja,entranament i aplicació del model.

Segona Part de la pràctica

La segona part de la pràctica consisteix en Visualitzar l'embedding amb l'eina TensorBoard de la suite tensorflow de google.

Per fer-ho caldrà adaptar al format requerit per tensorboard l'output de la fase 1 i carregar-lo al visualitzador de tensorboard.

Podeu instalar tensorboard portable o utilitzar una demo online al següent enllaç:
http://projector.tensorflow.org/

3 Consideracions i ajudes

Podeu trobar tota la documentació necessària als següents enllaços:

https://databricks.com/product/getting-started-guide

https://spark.apache.org/docs/2.2.0/mllib-feature-extraction.html

https://www.tensorflow.org/versions/r0.12/how_tos/embedding_viz/