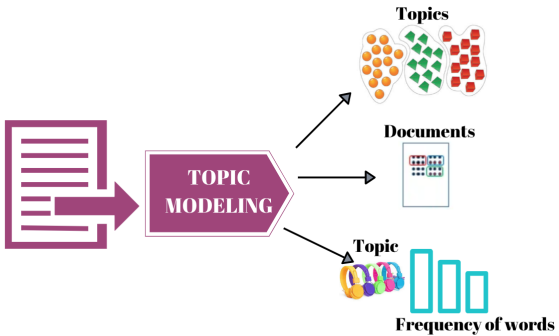
Top2Vec

Distributed Representations of Topics

Giovanni Gamaliel López Padilla

May 22, 2022

Introduction



Topic modeling is often used when a large collection of text cannot be reasonably read and sorted through by a person. Topic model will discover

- Latent semantic structure.
- Distribution of topics.
- Frecuency of topics.

Introduction

Traditional Topic Modeling Methods

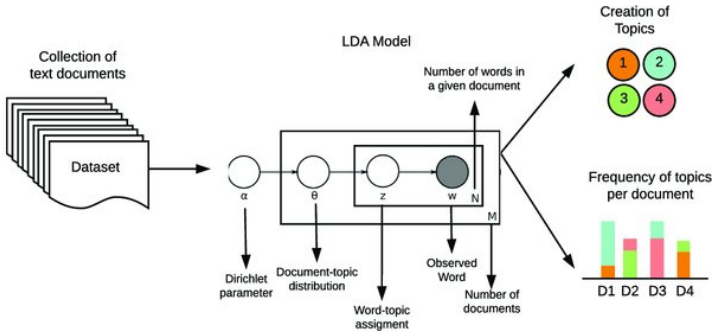


Figure 1: LDA representation [1].

Introduction

Traditional Topic Modeling Methods

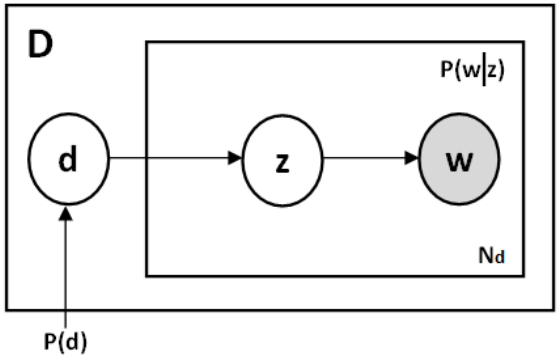


Figure 2: PLSA representation [2].

Introduction

Distributed Representations of Words and Documents

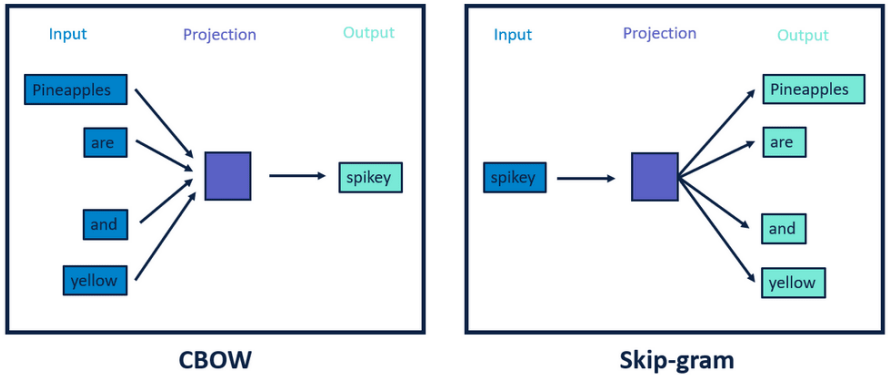


Figure 3: PLSA representation [2].

Introduction

Distributed Representations of Topics

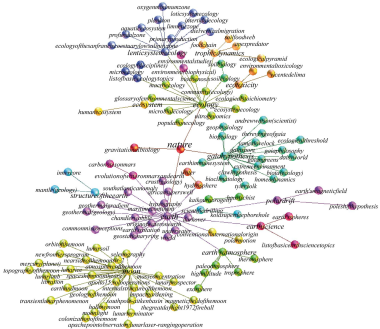


Figure 4: semantic space [3].

Model description

Create Semantic Embedding

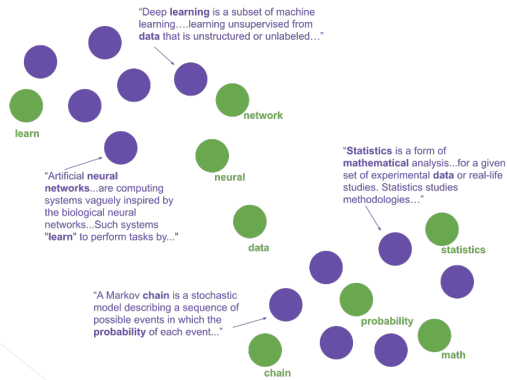


Figure 5: semantic space [4].

Model description

Find Number of Topics

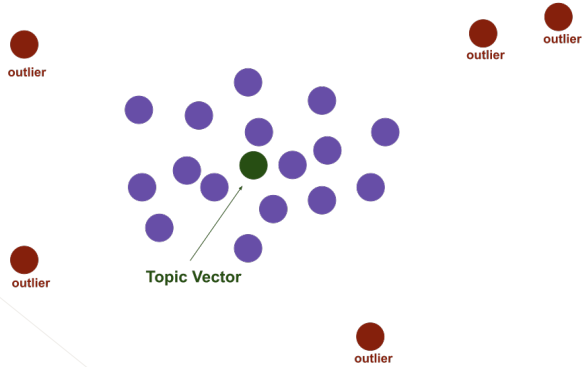


Figure 6: Topic subspace [4].

Model description

Low Dimensional Document Embedding

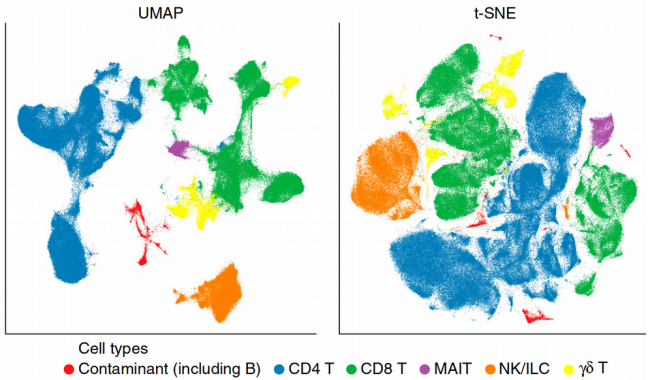


Figure 7: UMAP vs T-SNE [5].

Model description

Find Dense Clusters of Documents

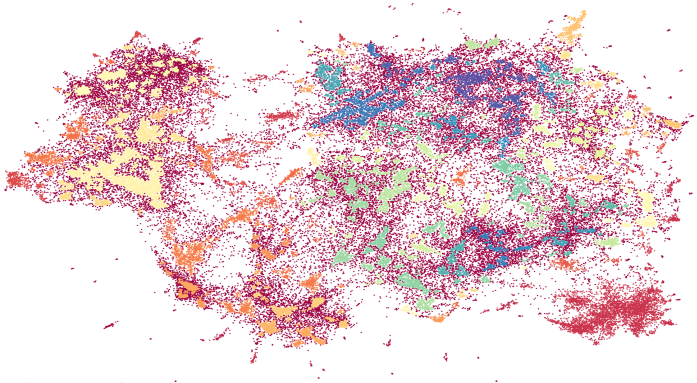


Figure 8: UMAP vs T-SNE [4].

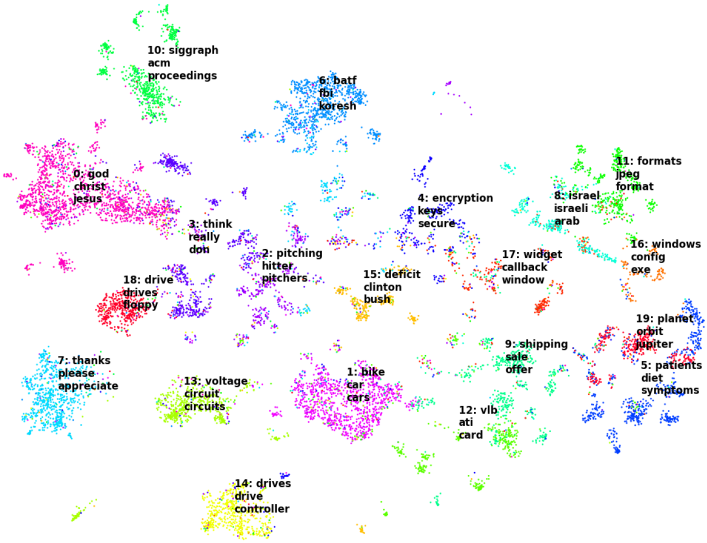


Figure 9: Semantic space for 20 News Dataset.

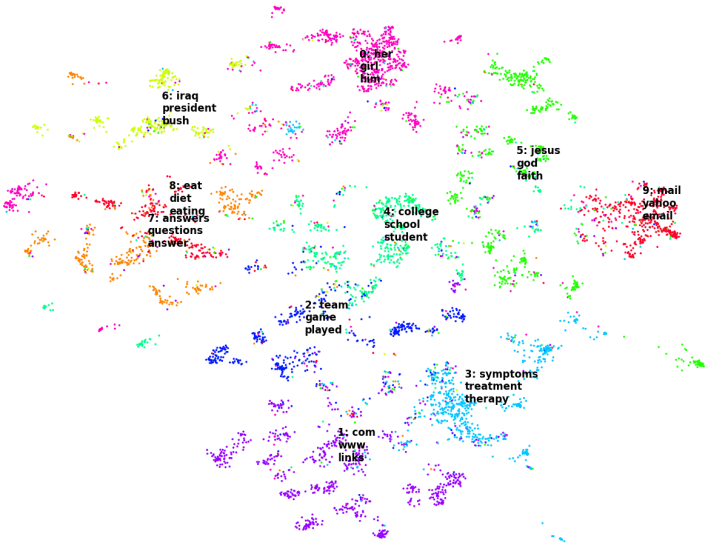


Figure 10: Semantic space for Yahoo Answers.

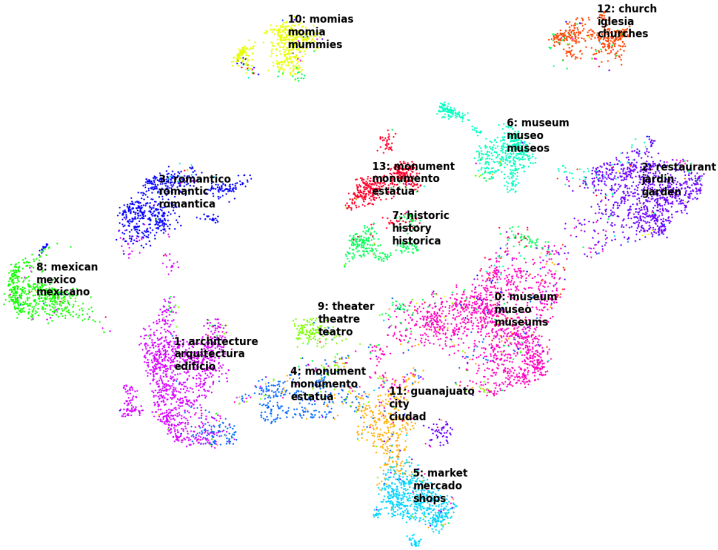


Figure 11: Semantic space for Tripadvisor.

Code

https://github.com/giovannilopez9808/Natural_Language_Processing_Proyecto

References I

Buenano-Fernandez D, Gonzalez M, Gil D, Lujan-Mora S.
Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach.
IEEE Access Access. 2020;8:35318–35330.
Available from: https://doi.org/10.1109%2Faccess.2020.2974983.

Alghamdi R, Alfalqi K.
A Survey of Topic Modeling in Text Mining.
ijacsa. 2015;6(1).
Available from: https://doi.org/10.14569%2Fijacsa.2015.060121.

References II

Masucci AP, Kalampokis A, Eguíluz VM, Hernández-García E.
Wikipedia Information Flow Analysis Reveals the Scale-Free Architecture of the Semantic Space.
PLoS ONE. 2011 feb;6(2):e17333.
Available from: https://doi.org/10.1371%2Fjournal.pone.0017333.

Angelov D.
Top2Vec: Distributed Representations of Topics.
CoRR. 2020;abs/2008.09470.
Available from: https://arxiv.org/abs/2008.09470.

References III

Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018 dec;37(1):38–44. Available from: https://doi.org/10.1038%2Fnbt.4314.