22/10/2019

Analyzing Hotel Reviews

Description:

Uncovering topics from texts has become an important task for many applications. Topic modelling provides methods to organize large collections of textual information. It can be seen as the unsupervised version of text classification. In order to accomplish this task, conventional models implicitly capture document-level word co-occurrence patterns to reveal topics [1].

Sentiment analysis is the computational study of people's appraisals and emotions toward entities, events and their attributes. The goal is to determine whether an opinionated document (e.g. reviews) or sentence expresses a positive or negative opinion. Opinions are important because whenever someone wants to make a decision, he would like to hear other's opinions [2].

TripAdvisor[1] is a famous website showcasing hotels with reviews. Users can write freely about their experiences with various hotels and also assign sub-ratings to specific aspects, such as cleanliness, location, or service. However, the number of hotel reviews can quickly be overwhelming, making it difficult for a user to find the information (s)he is looking for.

Currently, we have a large dataset of hotel reviews, annotated with the model of [3]: each word in a review has a probability distribution of belonging to a specific aspect (e.g. Cleanliness). By grouping phrases/sentences by masks we obtain multiple clusters, each corresponding to an aspect. In each cluster, many phrases/sentences are available (>100k).

The idea of this project would be to cluster phrases/sentences in each cluster and identify:

1) what they talk about (e.g. front-desk, cleaning service, waiters for the aspect Service);
2) the polarity of the mini-cluster;
3) if we could generate a summary.

Many methods could be used to achieve such a task: embedding based such as BERT [4] or Sent2Vec [5], key-phrase based [6], graph-based [7].

Prerequisites: Knowledge about Machine Learning & efficient with Python

Supervisor: Diego Antognini (diego.antognini@epfl.ch)

References:

[1] http://proceedings.mlr.press/v15/wang11a/wang11a.pdf
[2] https://www.cs.uic.edu/~liub/FBS/IEEE-Intell-Sentiment-Analysis.pdf
[3] https://arxiv.org/pdf/1909.11386.pdf
[4] https://arxiv.org/pdf/1810.04805.pdf
[5] https://arxiv.org/pdf/1703.02507.pdf
[6] https://arxiv.org/pdf/1801.04470.pdf
[7] https://www.aclweb.org/anthology/C10-1039.pdf

[1] https://www.tripadvisor.com/