OVERVIEW

At Crunchbase we collect news for different entities (e.g., an organization, an investor) that we track to power features such as event timelines, funding round discovery, daily digests, etc.

Our news dataset is composed of data from various publication sources. This data varies widely in content and style, as is natural when collecting news articles from the web. When collecting data, we face many engineering challenges such as content extraction, curation, inconsistencies, and volume.

The most interesting challenges involve tasks like identifying entities in the articles (or even better, identifying which entity the article is about!) or determining if the topic of the article is relevant to Crunchbase. Needless to say, this is a situation ripe for exploitation with data science techniques.

THE DATASET

The zipped plain text file (corpus.txt.zip) contains 10,000 documents in JSON format, where each document is newline delimited. The documents are the raw result of passing articles through the diffbot (https://www.diffbot.com/) analyzer we are using to extract content. Each document contains the raw text of the article along with additional metadata such as authors, mentioned entities, title, etc.

The corpus is sampled over time (by the time of when we fetched the article) and contains only English articles. However it is not otherwise normalized to any biases that may occur, such as volume per source, distribution of topics, length etc., rather it is a realistic sample of the data we collect.

THE TASK

Some of the articles in the dataset provided are not relevant to our use-cases. For example, some articles are concerned with tabloid celebrity gossip, restaurants, or sports events. At a high level, our goal is to identify the relevance of an article to the entities (e.g. an organization, an investor) tracked by Crunchbase.

While there are many different approaches to this problem, a first step might be to do some topic modeling to get closer to identifying what the article is about. **Your goal is to group the documents into cohesive but distinct topics.**

Please provide some documentation with your code and results that explains your approach along with its advantages and disadvantages. In addition, please provide comments on improvements or general next steps you might have taken had you more time to investigate and work with the dataset.

Hopefully you'll enjoy getting messy with some real world data :)

-Danielle