Automatic extraction of named entities and quantifying search relevance in news articles

Associated Press

Background

For obvious reasons, the Associated Press (AP) would like to be able to index and cross-reference its articles by topic (e.g. People, Places, Company, etc). This makes their data much more valuable to their clients, many of whom subscribe by topic in order to carry out applications in journalism, finance, and industry.

Problem Statement

Named Entity Extraction

The AP has a specific taxonomy of topics to which an article may be related. When a journalist submits an article for release, the AP uses a rule-based system to apply descriptive metadata (subjects, companies, organization, people, places) to tag its content using heuristics. There are several limitations of the current system:

- The rules are manually maintained, complicated to write, and difficult to maintain.
- They are generally based on simplistic rules like having a certain substring either present or not present.
- Ultimately, these rules these only imply the binary presence or absence of a topic rather than a relevance score (e.g. $\mathbb{R} \in [0, 1]$), so they can’t be sorted by presumptive relevance.

Search Relevance

Given a search term such as “Angela Merkel,” AP would like to automatically find and order search results which better reflect the degree to which a particular piece of content is about a particular topic. That is, is the given term not present, just a passing mention, or the central theme? Goals include:

- Benchmarking performance of the current search tools;
Finding methods and algorithms to get better classifications;
Gracefully handling slight misspellings using fuzzy matching;
A way to tune the relevance scores for high quality and consistent results for each topic.

Proposed solution

Research in the form of literate Jupyter notebooks that benchmark status quo performance and demonstrate new approaches for named entity extraction and search relevance.

Proof of concept in the form of a containerized software solution that implements the proposed solution, in the form of scripts for entity extraction and a REST API endpoint that implements searches on a partial corpus of AP articles.

Data Available

- Documentation describing our taxonomy and structure, as well as the API and data
  - https://developer.ap.org/ap-metadata-services
- Access to AP classification API, which can be used on arbitrary text
- AP News.com https://www.apnews.com/ search functionality
- For each story, AP News looks for string matches in the headline and top n characters to the value of classification terms on that story. This is not 100% accurate as it must match the term exactly, so for example the string 'Chris Wray' in an article will not show up because the classification term string is 'Christopher Wray'. It displays only those classification labels that match under 'Related Topics'. Other 'Related Topics' are added here from other sources such as particular feed names.
- Current classification rules for the Teragram system (in XML format) that apply AP taxonomy to content. AP writes and maintain these rules for Subjects, Organizations and Geography terms. Some People and Company terms have these rules as well for disambiguation purposes, but most are applied on a single canonical mention.
- At least 13,000 news text items that have appeared in APNews.com. The actual content scope is much wider than this – they have video, photo, audio, graphics and interactives content, both AP-sourced and from various other sources, that also has classification applied.
- (Possible extra) Access to New York Times taxonomy system and API for benchmarking current system