Meta-Analysis Towards Academic Consensus

Determine the degree of consensus in contentious academic fields

Capstone 2 Proposal

Jacques Poolman

In a recent debate between James Wilks and Chris Kresser on the Joe Rogan Experience[1] podcast, it could not be agreed upon what the current academic consensus is in many areas regarding plant-based diets. In a different field, for a while, many also argued divided academic consensus on climate change, and some still do.

While it seems that there is no official approach to determine academic consensus, at least in the medical field, no algorithms or guidelines exist to this end. Being able to capture and present key opposing views within an academic field as well as how they measure up against each other, will benefit through an automated meta-analysis. This could provide common ground from where to debate relevant issues and avoid wasting time on semantics.

Who will benefit from this data insight?

This problem might not have a clear business impact, but it could provide interesting information and insights on the degree of consensus in certain academic fields.

An automated tool could analyse any keyword and provide a quick reference for researchers, news reporters interested in specific academic or scientific fields, as well as new technologies. A rapid meta-analysis could have further-reaching applications.

Data

In considering various APIs, this model will use the Public Library of Science (PLOS) API since it also includes searches by Abstract. In addition, according to the PLOS API Documentation[2], the API does not require an API Key for access to almost all works, with the exception of sources that don't allow redistribution of data. An API Key would otherwise only be required to add or update works.

---

[1] http://podcasts.joerogan.net/podcasts/james-wilks-chris-kresser-gamechangers-debate
[2] http://alm.plos.org/users/me

The ‘requests’ Python library will request the documents from the API in batches of 100 (maximum allowed) and append it to a dataframe called 'corpus'. A 10 second delay will be built in between requests to comply with PLOS API rate limit.

Method

Inspired by the Joe Rogan Experience podcast mentioned above, this project will explore the somewhat contentious academic field of nutrition. To this end the *academic_consensus* model will search the abstracts of academic papers that contain the keyword "nutrition".

NLP models will identify and categorise concepts in this field and determine statistical significance between opposing 'truths', if any. Ranking these groups according to weighted influence will prove the degree of consensus of various approaches in a given academic field.

Following Patrick Harrison’s “Modern NLP in Python” at PyData DC 2016 as a guideline, the analytical approach wil include:

- Phrase modelling
- Topic modelling with LDA
- Visualising topic models with pyLDAvis
- Word vector models with word2vec
- Visualising word2vec with t-SNE

Productionise

Deliverables

Report, webpage and Jupyter Notebook code on GitHub.