Technical Screen - NLP

Please answer as best as you can, and send your answers back in a google folder.

1. Answering Questions

Please give a detailed explanation about how you would generate a model for answering questions like the one listed below, take from CS224N:

Which team won superbowl 50?

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24-10 to earn their third super bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

2. Communicate For Me

We want to generate observations from abstracts to prospective clients. For an article like this, here's an example of an email we would send to a prospective collaborator.

> Dr. Tannir,
>
> I read your recent article on Bempeg + nivolumab. Strong ORR performance; for the patients who achieved complete response, did you do any immune profiling?
>
> I'm a scientist with teiko.bio; we've developed a simple, rapid service on top of mass cytometry to comprehensively profile immune cell responses from peripheral blood. You may be interested in our cofounder's Cell paper, "Comprehensive Immune Monitoring of Clinical Trials to Advance Human Immunotherapy."
>
> Would this be interesting for your work?
>
> -Bill Kapri | Book a time

Please provide an explanation of how you would go about building a model to accomplish this task. If it helps to be specific, here is an example article and the email template is below. The key parts we want to generate are:

- SHORT_DESCRIPTION

- OBSERVATION_ABOUT_RESEARCH
- DESCRIPTION_OF_STUDIES.

  [Author first name],

  I read your recent article on [SHORT_DESCRIPTION]. It's nice to see [OBSERVATION_ABOUT_RESEARCH]. How are you characterizing immune response in your [DESCRIPTION_OF_STUDIES]?

  I'm a scientist with teiko.bio; we've developed a simple, rapid service on top of mass cytometry to comprehensively profile immune cell responses from peripheral blood. You may be interested in our cofounder's Cell paper, "Comprehensive Immune Monitoring of Clinical Trials to Advance Human Immunotherapy."

  Would this be interesting for your work?

  -Bill Kapri | Book a time

I would consider using techniques including Gensim, LDA model, TF-IDF, and BERT to do topic modeling. [SHORT_DESCRIPTION] is likely to be found using the abstract or introduction. [OBSERVATION_ABOUT_RESEARCH] may be discovered with the conclusion. [DESCRIPTION_OF_STUDIES] can be identified with the introduction, literature review, or background.

3. Free Response

Please give a detailed description of how to ensure that the NLP code that you generate so that it can be shared with and operated by others to work on different datasets or to build on top of the work you have done?

Furthermore, I will generate a Readme file at the end of the internship.

4. Web scraping

Take the following link to a journal and return a csv list of titles, authors, affiliations, and URL. Please return your runable code and the csv outputs in a google drive folder.

Questions 5 and 6 are taken from the following dataset used for sentiment analysis.

5. Computation

Using the following data set, please perform preprocessing of the data:

1. getting rid of non-alphabet characters
2. converting to lowercase
3. removing stop words
4. performing lemmanization

6. Modeling

Using the data set above, train a random classifier model, and evaluate the model returning metrics such as:

1. Accuracy Score – no. of correctly classified instances/total no. of instances
2. Precision Score – the ratio of correctly predicted instances over total positive instances
3. Classification Report – report of precision, recall and f1 score