Coursera Transcripts Analysis
-Topic Modeling

By Dajun Lin, Natalie So, Amy Zhao

Purpose of the project

- Analyze Coursera transcripts

- Extract a list of main topics

Why we find it interesting?

Beneficial to User:

- To analyze the main knowledge points of the course.
- To capture the similar knowledge points in the different courses that the user has learned

Beneficial to Coursera Administrator:

- Categorize courses
- Make recommendations to user

Workflow of the Code

1. Text cleaning: text content retrieval; root word
2. Topic analysis by Latent Dirichlet Allocation (LDA) model
3. Display of interactive chart by pyLDAvis

Libraries Used

- Pandas
- Numpy
- Spacy
- Gensim
- Nltk
- pyLDAvis
- Streamlit

Text Cleaning

```
def tokenize(text):
    lda_tokens = []
    tokens = parser(text)
    for token in tokens:
        if token.orth_.isspace():
            continue
        elif token.like_url:
            lda_tokens.append('URL')
        elif token.orth_.startswith('@'):
            lda_tokens.append('SCREEN_NAME')
        else:
            lda_tokens.append(token.lower_)
    return lda_tokens
```

```
def get_lemma(word):
    lemma = wn.morphy(word)
    if lemma is None:
        return word
    else:
        return lemma


def get_lemma2(word):
    return
WordNetLemmatizer().lemmatize(wo
rd)
```

Topic Analysis by LDA

```
pickle.dump(corpus, open('corpus.pkl', 'wb'))
dictionary.save('dictionary.gensim')


NUM_TOPICS = 5
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics=NUM_TOPICS, id2word=dictionary, passes=15)
ldamodel.save('model5.gensim')
topics = ldamodel.print_topics(num_words=4)
```

Display of interactive chart by pyLDAvis

```
vis = gensimvis.prepare(ldamodel, corpus, dictionary)
html_string = pyLDAvis.prepared_data_to_html(vis)
st.components.v1.html(html_string, width=1300, height=800)
```

Application deployment

We use Streamlit for a web app visualization

https://coursera.streamlit.app/

Thank you!