Sayed Ahmed, Toronto, Canada
Linkedin: SayedJustetc

8112223 Canada Inc
JustEtc Social Services

CSE, BUET

ShopForSoul.com
Training.SitesTree.com
Bangla.SaLearningSchool.com
Youtube: SaLearningSchool-ShopForSoul

BSc. Eng. in Computer Sc. & Eng.
MSc in Computer Science
MSc in Data Science and Analytics

Workplace Communication Program
Teach in Higher Education

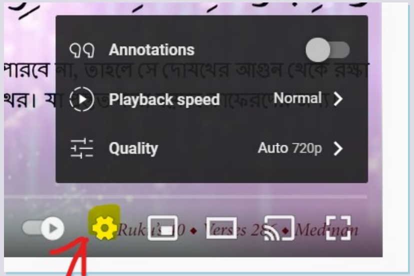
Linkedin Learning
IBM Data Science/CognitiveAI
SkillSoft

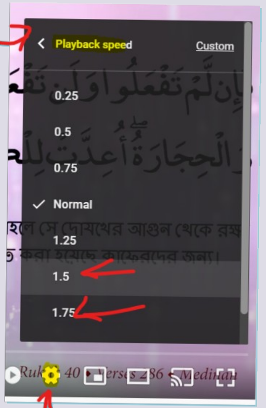
NLP – Natural Language Processing

Ref: Internet Images

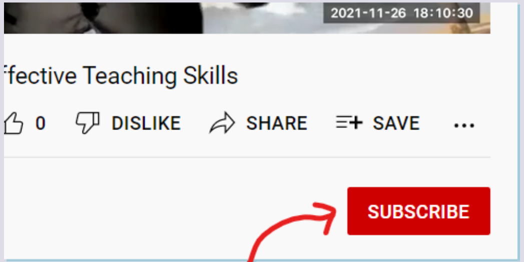
Youtube
Listen Faster/Slower

- Watch at 2x speed or 1.5x speed





- Subscribe



Misc

- Buy the courses
  - https://ShopForSoul.com
  - http://sitestree.com/training/
  - http://bangla.salearningschool.com/
- Get access to our LMS

- Advantages
  - Discussion
  - Chat
  - Live Sessions
  - Select topics to create videos on
  - Q & A
  - Free Courses

NLP Problems (Assignments) to Solve

- Regular expression to extract data and information from text
- Tokenization using NLTK or similar
  - Word token
  - Sentence tokenization
  - Utilize regular expression
- Lemmatization using NLTK
- Stemming using NLTK or similar
- Remove stopwords from text
  - Then tokenize
  - Do lemmatization and stemming after stop word removal

NLP Problems (Assignments) to Solve

- NLP: Write code to remove punctuations from text.
- NLTK: Take stop words list from library. Add your own stop words to the list
  - Then remove stop words from a text
- Write N-Gram (Bi Gram, Trigram) code using frequence as the measure
- Write N-Gram (Bi Gram, Trigram) code using collocated words as the measure
  - What are collation words
  - Find a library method or a 3rd party implementation, use it as well
  - Compare your output with the library/3rd party one
  - Print top few N-grams

NLP Problems (Assignments) to Solve

- Find about NLTK methods/features such as
  - ngram_fd
  - ngram_fd.items()
  - dir(trigrams)
  - help(trigrams.ngram_fd)
  - nbest
  - TrigramAssocMeasures
    - raw_freq
    - help(TrigramAssocMeasures)

- NLTK features for
  - MLE (Maximum Likelihood Estimate) and Laplace smoothing
- Implement
  - trigrams based smoothing using Laplace and Kneser Ney algorthms
- Implement
  - Laplace smoothing
    - Bigram, trigram
  - Measure the preplexity
  - Measure prepexity as
    - a) Logbase2Prob= Sum-for-all-trigrams(log2(P(w3|w1,w2)) (b)Ent=(-1/tokens-in-test)x Logbase2Prob (c) Power(2, (ent))

NLP Problems (Assignments) to Solve

- Write a program to predict the next few words
  - Based on bi-gram and tri-gram
  - Based on a sample text
  - Use train and test approach
- Study this implementation
  - https://github.com/smilli/kneser-ney
  - Utilize utenberg corpus is required
-

NLP Problems (Assignments) to Solve

- Using Penn Treebank, do POS tagging to a text
  - http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- Find out what are these
  - Treebank corpus and Brown corpus
  - Can you use these for the POS tagging task above
- Write short essay on
  - Chunk grammar
  - http://www.nltk.org/book/ch07.html.

NLP Problems (Assignments) to Solve

- From some example text
  - extract three different chunks of your choice: e.g., co-ocurrences of adjectives and nouns, co-occurences of determiner, adjectives and nouns, extractions of all types of nouns, etc.
  - Learn on how to train a custom tagger at
    - http://www.nltk.org/book/ch05.htm
- Train an HMM model on the sentences of Brown corpus. Find out the accuracy of your trained HMM model on the sentences in test data

NLP Problems (Assignments) to Solve

- Use NLTK's tagger to predict the tags and determine accuracy of prediction.
- Read on
  - Maximum Entropy Classifier (MaxEnt)
  - Why MaxEnt is highly accurate
- Read on a decision tree classifier for POS tagging.
  - http://nlpforhackers.io/training-pos-tagger/

NLP Problems (Assignments) to Solve

- Classify text using
  - Naive Bayes
- Use movie data from here and classify the reviwes to be positive or negative. Use train and test
  - http://ai.stanford.edu/~amaas/data/sentiment/
  - load_files scikit-learn
  - CountVectorizer
- Modify the above implementation
  - Get rid of the words occuring in more than 1000 documents
- Redo the above implementation after modifying the text
  - Such as add Not when you see a negative word, and till the first punctuation

NLP Problems (Assignments) to Solve

- Use the sentiment lexicon below
  - http://sentiment.christopherpotts.net/lexicons.html
  - Create two features
    - Positive words count
    - Negative words count
- Filter out some words using POS tagging
  - Keep adjectives, verbs, and nouns
  - And then try the sentiment analysis
  - i.e. text classification

NLP Problems (Assignments) to Solve

- Try this example on Neural Network and text classification
  - https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/ (another beginner tutorial)
  - Try to improve the accuracy
- Train a Neural network with embedding
  - Use the IMDB database as above
  - For text classification
  - Measure the performance and compare the result with Naive Bayes

NLP Problems (Assignments) to Solve

- Repeat the text classification after replacing negative words wit
  - NOT
- Glove: https://nlp.stanford.edu/projects/glove/
- Pretrained word to dense vectors
- Use glove for a multi-label classification problem
- Use glove and build a binary classifier
  - Pick one of the toxicity columns as the class
  - Classification for Wikipedia comments

NLP Problems (Assignments) to Solve

- “Make a binary classifier for each class, and assign multiple labels (classes) to each test record. Evalute your accuracy for multiple classes. This is little more work but is more rewarding as a learning experience.”
  - Use Glove
  - And wikipedia comments

NLP Problems (Assignments) to Solve

- Try the name entiry recognition information and example
  - https://www.depends-on-the-definition.com/sequence-tagging-lstm-crf/
  - https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus/version/4#ner_dataset.csv
- Create a Bi-directional LSTM (RNN)
  - For Name Entity Recognition
- Modify the NER example using
  - Glove
- Concatenate each word with POS tagging
  - And then adjust the LSTM/RNN for NER

NLP Problems (Assignments) to Solve

- Implement Knee/Elbow method of text clustering
- Determine purity of clusters
- Utilize Bernoulli mixture model of clustering
- Explain the Bernoulli mixture model of clustering model as can be seen in
  - https://github.com/manfredzab/bernoulli-mixture-models
  - https://github.com/schwannden/MNIST_mixture-of-bernoulli

NLP Problems (Assignments) to Solve

- Gaussian Mixture models of clustering
  - http://scikit-learn.org/stable/modules/mixture.html#mixture
- Implement LDA topic modeling algorithm
  - Utilize train/test
- Implement PLSA topic modeling algorithm
- Utilize PLSA topic modeling algorithm from Scikit-learn and apply on a dataset
-

NLP Problems (Assignments) to Solve

- Read the Topic Modeling blog at
  - https://nlpforhackers.io/topic-modeling/
- Implement
  - CountVectorizer (Freq)
- Take a set of text/articles/news
  - Train LDA on the data and find the top topics
  - Apply EM if applicable
- Implement PLSA topic modeling algorithm
  - TruncatedSVD
- Modify the code in the URL to get rid of noise from the tokens. Remove *,/,-,=,,_ or similar

NLP Problems (Assignments) to Solve

- Implement LDA and PLSA
  - And apply on some Gutenberg project data
  - Find the topics
- Use LDA, PLSA to find topics
  - Use these topics as features for Naive Bayes
  - Then implement a Naive Bayes classifier
  - Find: accuracy, precision and recall

NLP Problems (Assignments) to Solve

- Read on Gensim
  - https://radimrehurek.com/gensim/models/ldamodel.html
  - Implement LDA
  - Use alpha and beta
  - Use Gensim

NLP Problems (Assignments) to Solve

- Implement Textrank algorithm
  - Use Gensim
- Take articles from the Internet
  - Implement text summarization
  - Implement keyword extraction using Gensim
- Implement Naive Bayes classifier as below
  - Find 100 key phrases/keywords from each document/review
  - Merge these keywords and
  - Filter these keywords from the documents
  - Then with the remaining data train and implement Naive Bayes
    - Calculate accuracy, precision, and recall
- Implement ROUGE metric for text summarization
  - “ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing” Wikipedia