News Headline Generation

04.07.2018

Final Project:

INFO 7374 Cognitive Computing and Deep Neural Networks

Naini Shah

Vighnesh Venkatakrishnan

Overview

Newspaper plays a significant role in our day to day life. In a news article, readers are attracted towards headline. Headline creation is very important while preparing news. Our goal is to implement text summarization by generating headline for a news body using recurrent neural networks. We attempt to reproduce the results in the paper: Generating News Headlines with Recurrent Neural Networks

Goals

1) To build a web app prototype that generates headline for a given article using a model trained with variants of RNN (e.g. LSTM) in Keras with gigaword dataset.
2) Attempt to find similarity between actual headline and generated headline using cosine similarity
3) Try to train the model with data scraped from the internet as suggested in the paper referenced below

Bonus goals (if time permits)

1) Attempt to train the model using other frameworks like PyTorch and compare the generated outputs
2) Implement news headline generation for a language other than English
3) Try ensembling the deep learning models and check if there is improvement in results

Use Cases

1) **Headline generation helper:** Sometimes it is tricky for a person to think of a nice headline, our app could assist the user with that.
2) **Attempt to curb click baits:** Sometimes there are misleading links which claim to talk about some news but end up describing something else. Our app can prompt the user before clicking such links.

Datasets

1. English Gigaword: https://catalog.ldc.upenn.edu/LDC2003T05
2. Google News dataset
3. Data scraped from reliable news sources

The paper uses English gigaword dataset. We plan to use combination of gigaword and google news.

Process Outline

1. Data Preprocessing
    - Data Cleaning, handling missing values
    - Join the gigaword data and google news data
    - Web scaping data (later)
2. Splitting the cleansed data into training and test set
3. Building RNN model
4. Generate predicted headline
5. Build a web app prototype to demonstrate the same
6. Deploy web app in AWS (future scope)

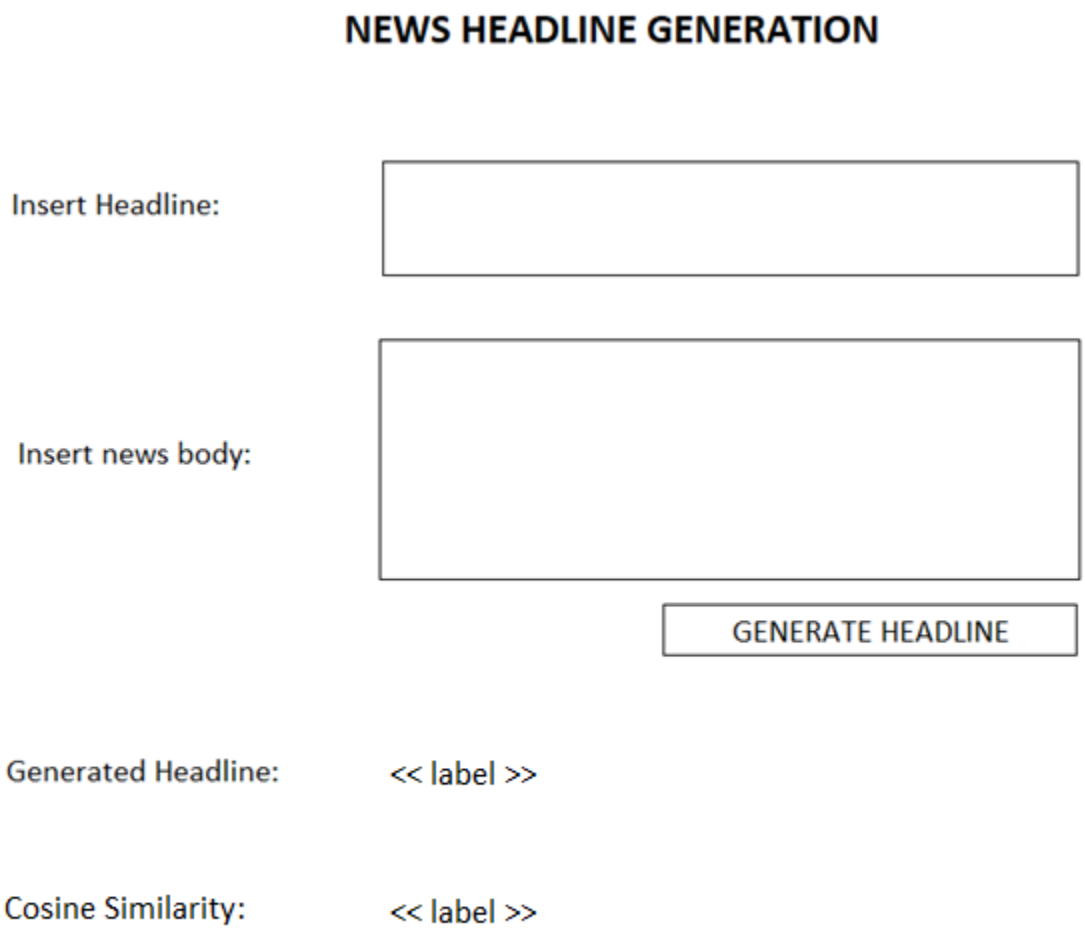
Milestones

Timeframe	Delivery
Week 1 (31st March–6th April)	Project proposal presentation
Week 2a (7th– 9th April) Week 2b (10th-13th April)	Finalize project proposal Data preprocessing
Week 3 (14th-20th April)	Build/Train model and keep web app prototype ready
Week 4 (21st-27th April)	Implement cosine similarity and work on future scope if time permits

Deployment Details:

1) Language: Python 3.6
2) Cloud Tools/Platforms: AWS EC2
3) Development Environments: Jupyter Notebook
4) Packages/ Frameworks: Keras

User Interface Design Plan



Reference and Sources:

1. https://arxiv.org/pdf/1512.01712.pdf
2. https://github.com/fchollet/deep-learning-with-python-notebooks/blob/master/8.1-text-generation-with-lstm.ipynb
3. https://github.com/udibr/headlines
4. https://github.com/kabrapratik28/DeepNews