The Office:

Modeling Iconic Character Arcs

Abstract

This project seeks to model the character arcs of main characters from The Office using natural language processing and both semi-supervised and unsupervised topic modeling algorithms. After tuning models, I finalized CorEx topic models with 3-6 topics for Michael Scott, Dwight Schrute, Pam Beesley, and Jim Halpert and presented visualizations of the topics' prevalence in each character's storyline over 5 seasons.

Design

This binary classification model predicts instance of diabetes based on a person's baseline health characteristics, including Blood Pressure, Glucose Levels (two hours after consuming a sugary drink), Age, and BMI. The data were trained on multiple models, the two top-performing models were tuned using Random Search CV, and the final model was selected via K-fold cross-validation, and the scoring metric used was F2-Score and Recall. These metrics were chosen to ensure a minimal number of false negatives, as false negative diagnosis of diabetes could prove extremely dangerous to the health of patients.

Data & Algorithms

The dataset contains scripts from 66 episodes.

- Scraping scripts from ForeverDreaming
    - ✳ Randomly selected 50% of episodes from each season, for a total of 66 episodes.
    - ✳ Created separate data frames for each character, with each row containing the character's lines from an episode.
- Preprocessing using regex and SpaCy
    - ✳ Randomly selected 50% of episodes from each season, for a total of 66 episodes.
    - ✳ Regex to remove punctuation, bracketed text, bracketed directions, i.e. [laughs], [drops bucket of chili].
    - ✳ SpaCy removed verbs and adjectives, transformed all tokens into their lemmatized form.

    - ✳ Using the NLTK English stop word list as a base, I created a customized stop word list for each character. I accounted for nicknames by swapping out nickname for full character name, i.e. for Dwight, ‘Monkey’ was replaced by ‘Angela,’ for Michael, ‘Pammy’ and ‘Spamster’ were replaced by ‘Pam.’ Ordinal and time-based, work-specific nouns were removed due to high frequency, i.e. ‘Morning’, ‘Weekend’, ‘Tuesday’, ‘Month’, ’Desk’, ‘Stapler’, ‘Computer’, ‘Cell.’
- I fit baseline vectorizers (count vectorizer and term frequency inverse document frequency) and tuned parameters
    - ✳ I find for most supporting characters, a high max_df (0.7-0.85) is preferable, whereas a smaller value was preferable for Michael, who has significantly more lines than any other character. min_df and max features were
- I fitted unsupervised & semi-supervised topic modeling algorithms
    - ✳ PCA, SVD, LDA, NMF, usually fitting between 3-10 topics
    - ✳ Used the most prominent word groups from unsupervised models, plus domain knowledge of the episodes to inform anchored topics used in CorEx models.
- Build Visualizations & Present Findings
    - ✳ Using soft prediction outputs from each CorEx model to map changes in topic themes over the seasons
    - ✳ Identify most prominent character relationships

Tools

- Numpy, Beautiful Soup & Pandas for Data Scraping and compilation
- SpaCy, regex, and NLTK for the preprocessing pipeline
- Scikit Learn, Gensim for vectorization and unsupervised semantic analysis
- CorEx for semi-supervised analysis,.
- Seaborne, Scattertext, and MatPlotLib for visualizations