Andrew Sutardji
December 9, 2016
CSCI 4243W

Design Document

Purpose

SmartCaster is an automatic storytelling program, actualizing natural language processing and machine learning to generate coherent texts from subject inputs. In its current form, it constructs brief multi-sentence articles from weekly NFL statistics. The information relayed to the user is of their choice, allowing them to only receive updates on the players and teams they care about. As opposed to the sports columnist who spends unnecessary time putting together summaries and highlights at the end of games, *SmartCaster* looks to put out content of similar, if not better, quality almost instantly. Concerning analytics and statistical evaluation, *SmartCaster* strives to be free of the human biases that naturally make their way into composition, especially in a realm of entertainment that manages to stir passions within American culture.

SmartCaster is purely informative at its core, but language may always be catered to fit the user. A dictionary and thesaurus that can recognize the "degree" of adjectives and their synonyms may also be used to generate biases by putting positive and negative emphases on certain subjects, based on the user's preferences. This personalized and regulated practice of injecting sentiments into these customized articles creates a sense of collusion between the user and the publisher, promoting a relationship of reinforcement and validation.

The ultimate goal of *SmartCaster* is to explore human understanding of the English language. Encoding grammatical syntax and strictly defining the proper use of words solidifies the logical structure of English. Stemming from the belief that technology exists to make human life more convenient, automated storytelling programs like *SmartCaster* eliminates the thoughtless practice of translating hard data and fact into English syntax.

Audience

The primary intended audience for *SmartCaster* consists of American football fans. The approach of overstimulation on behalf of the entertainment industry, especially within the sports industry, does enough to push perspectives and biases upon the viewer. Those readers typically take away the emotional appeal of the author, rather than an actual understanding of the information. *SmartCaster* wishes to deliver the hard data through simple language that doesn't conflict with the user's own perspectives.

Another major target audience *SmartCaster* strives to appeal to is fantasy football team owners. The analysis provided by current fantasy football applications, hosted by ESPN and NFL, rely on generalizations based on aggregated data over the course of a season. The summaries published per player are insufficient in properly evaluating their performance to these team owners, sometimes delivering bad advice on statistics reviewed through biased perspectives.

A possible audience for *SmartCaster* is football analysts. Money and sports go hand in hand, and careers can depend upon thoughtful evaluation of the available data. *SmartCaster*, with a continually developing content determination system, can assist in taking a player or a team's history and providing possible mappings of factors and their performance.

Use Cases

Since the sole service of *SmartCaster* is to relay NFL player, team, and game summaries to the user, it would be the only objective of the program's use. The marketed, innovative element of this delivery would come through natural language generation. Small content, such as brief summaries, would be the most readily accessible material to the user.

The content the user sees would come according their selected preferences to minimalize the effort of parsing through unsolicited information. In minimalist fashion, the user is only presented with features or content that they explicitly request, whether it be switching between categories or reformatting their own preferences.

Comprehensive tables, or charts, of data is another option for viewing data. Although this eliminates the need for the natural language aspect, it is a simpler matter providing graphical representation of data than through linguistic means.

Major Components

Data Retrieval

The NFL API, written in Python, gathers data weekly from NFL Game Center and makes player, team, and game statistics available as JSON data. From these objects, specific data can be retrieved via attributes, contributing to a temporary corpus of only necessary information. If specified information is not available within certain contexts, the program should recognize the problem and proceed if it is seen as an anomaly.

Estimated time of completion: no more than 1 week.

Data Analysis

Once relevant data is aggregated, it must be evaluated. This comes in the form of averages, charting, trend recognition, and highlighting drastic points of change. The Python Data Analysis Library (pandas) has the means of parsing through this data and doing the tedious, menial labor of looking through all this simple information.

Estimated time of completion: ~4 weeks. May be done in tandem with content determination. Will require continual updating and correction of algorithm.

Dictionary & Thesaurus

In striving for the most proper way to compose text, *SmartCaster* must abide by English syntax. This means following the set of grammatical rules and tenses already rigidly defined in the world of authorship. The Natural Language Toolkit (NLTK) has its own method of conveniently tokenizing texts and identifying the parts of speech of words in their context. WordNet functions as a thesaurus, lending a hand in diversifying word usage in texts.

Estimated time of completion: ~4 weeks. Will be an ongoing process as the dictionary may be expanded to appeal to more markets.

Content Determination

Providing the user with concise and non-imposing content means determining what information is the most pertinent to the user. The need to weigh the importance of averages and highlights comes into question here, and will be determined by the frequency of some cases, including positive and negative trends and relevance of an event within the context of a game. This aspect will be built with pandas, but likely will have to find many alternative routes.

Estimated time of completion: ~6 weeks. May be done in tandem with data analysis. Will require continual updating and correction of algorithm.

Textual Composition

Parse trees will be the most effective structure to ensure the composition's adherence to English syntax. Markov chains would also come into play when dictating sentence diversity throughout the text. The thesaurus will factor in when imbuing the content with a desired sentiment, and the dictionary will weigh in against any minor changes to ensure grammatical correctness.

Estimated time of completion: ~8 weeks. Will be done alongside content determination. Will require continual correction of algorithm to ensure sentence diversity, as well as the inclusion of further audiences and wider ranges of rhetoric.

Textual Review

Entity extraction programs, available through NLTK and AlchemyLanguage, provide "scores" as to how relevant any given subject within the text is against its context. AlchemyLanguage also provides sentiment detection, useful in reviewing the nearly finished product before the reader receives it. These quantified "scores" can be evaluated as parameters, requiring a threshold for specified qualities before publication.

Estimated time of completion: ~2 weeks. Will be done alongside textual composition.