Project Scope document – Russell Conjugation

DATA SCIENCE PRACTICUM

PRAGYA MISHRA – A20406346

Contents

Project Scope Document

1   Project Name

Research project on Russell Conjugation

1.1   Location of Document

https://docs.google.com/document/d/1z3FORu4oL5Q9LMcyNjVhZEDoENgjFQBSd_8twJLdS0w/edit?usp=sharing

2   Team Composition

2.1   Sponsor:

Prof. Shlomo Argamon

2.2   Illinois Tech:

Prof. Shlomo Argamon

3   Project Background

3.1   Sponsor's Line of Business:

Research in computational methods for style-based analysis of natural language using machine learning and shallow lexical semantic representations, exploring application in intelligence analysis, forensic linguistics, biomedical informatics, and humanities scholarship.

3.2   Description of Sponsor's Problem:

The use of subtly emotive language can bias interpretation of otherwise objective and accurate characterizations of people and events. Speechwriters and rhetoricians have used careful word choice to good effect since time immemorial. Bertrand Russell memorably encapsulated the idea in pseudo-conjugations such as:

- I am firm, you are obstinate, he is a pig-headed fool.
- I am righteously indignant, you are annoyed, he is making a fuss over nothing.
- I have reconsidered the matter, you have changed your mind, he has gone back on his word.

Here, pairs of words such as 'firm', 'obstinate' and 'pigheaded' are known as Russell or Emotive Conjugates of each other. As rhetoric in all media, both political and non-, has become increasingly polarized, so has, it seems, the use of such emotive language to pre-emptively destroy one's opponents and prop-up one's heroes.

4   Project Objectives and/or Goals:

To build and deploy a prototype bias-revealing browser plugin, which will reveal hidden sources of emotive bias (Russell Conjugates) in online rhetoric. The long-term goal of this project is to investigate to the extent to which Russell conjugations are used to bias rhetoric, to develop tools to make readers aware of such rhetorical tricks, and to investigate how such tools affect readers' perceptions of bias and evaluation of information. The scope of the current project is to build a browser plug-in that automatically

a) Identifies source of polarization in a rhetoric, i.e., emotively connotated words,
b) identify and present Russell Conjugates of these emotive words as an option to the user.

5   Project Deliverables

5.1   Type of Deliverables

Required:

a) A model that classifies pairs of words as Russell Conjugates or otherwise.
b) A dataset for Model building

Optional:

A browser plug-in that
    a. Identifies source of polarization in a rhetoric, i.e., emotively connotated words,
    b. Present Russell Conjugates of these emotive words as an option to the user.

5.2   Expectation on Deliverables

a) Classification model should have a good cross-validated performance, i.e., F1 score > 0.8 and Accuracy > 85%
b) The dataset must have both positive and negative examples of Russell Conjugates.

6   Project Risks and Constraints

6.1   Risk

a) The data generated might not be enough.
b) Since the data will be created manually, human errors can be present in building the dataset.
c) Prediction of the model will not be 100% correct.

6.2   Constraint

a) No prior computational work on Russell Conjugates. Thus, there is lack of references.
b) No prior dataset present to be used as reference.
c) Dependency on third party packages like Scikit-learn, Matplotlib etc.
d) Time spent per week on the project is approximately 30 hours.
e) Applicable to only English language.
f) Dependency on pre-trained word vectors by Google using word2vec

7   Project Execution

7.1   Strategy: Agile Method

7.2   Agile Development: Epic, Story, Task

Epic is a sub-project, Story is a collection of related tasks, and Task is the tangible work that you are going to complete.

Epics	Stories	Tasks
Phase 1	Literature survey	Literature survey on Computational work on emotive conjugations
		Literature survey on media bias, its type and influence
		Literature survey of Word Embeddings
	Dataset creation	Collecting examples of Russell Conjugates by comparing different news coverage on the same topic
		Filtering the examples to create positive examples of pairs of Russell Conjugates:
		1. Ensuring pairs of words to have same POS and participle.
		2. Include same of pairs of positive examples but with their order switched (for example firm, obstinate and then obstinate, firm)
		Creating negative examples by randomly combining words from dictionary

		Keeping the ratio of positive to negative example as 1:10 to increase variety of negative examples
Phase 2	Data Preparation	Using pretrained word vectors using Google's word2vec
		Converting each word pair from data into a difference vector using the pre-trained vectors
	Modeling and Validation	Creating baseline linear classification models
		Using oversampling to reduce bias due to unequal proportion of positive and negative examples
		Performing dimensionality reduction using PCA to get a sense of highest variant directions for each class
		Building classifier 1 - identifying contexually & denotationally similar pair of words
		Validation and fine-tuning classifier 1
		Building classifier 2- filtering out synonyms
		Validation and fine-tuning classifier 2
		Research other approaches to identify/generate Russell Conjugates
Phase 3	Creating Russell Conjugate database	Decide a thesaurus to be used
		Use the model to classify pairs of words in thesaurus to build a database
Phase 4	Web Plug-in (Optional)	Create a module that analyses a web page's text to highlight emotionally connotated words
		Use the database of Russell Conjugates created earlier to provide users with alternate words
		Deploy the above two functionalities as a Web Plug-in

7.3   Timeline

Stories	1/17	1/24	1/31	2/7	2/14	2/21	2/28	3/7	3/14	3/21	3/28	4/4	4/11	4/18	4/25	5/2
Literature survey	■	■	■													
Dataset creation		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Data Preparation				■	■											
Modeling and Validation					■	■	■	■	■	■	■	■	■	■	■	■
Creating RC database using model												■	■			
Web Plug-in (Optional)												■	■	■	■	■

8   Document Revision History

3/15/2019 – Pragya – Initial Version

5/3/2019 -Pragya