Linguine Requirements Document

Team Rigatoni

# Introduction

* 1. Document Purpose
  2. Intended Audience and Reading Suggestions
  3. Vision and Scope
     1. Vision Statement
     2. Major Features
     3. Assumptions and Dependencies
     4. Scope of initial increment
     5. Scope of future iterations
     6. Scope beyond this year’s project

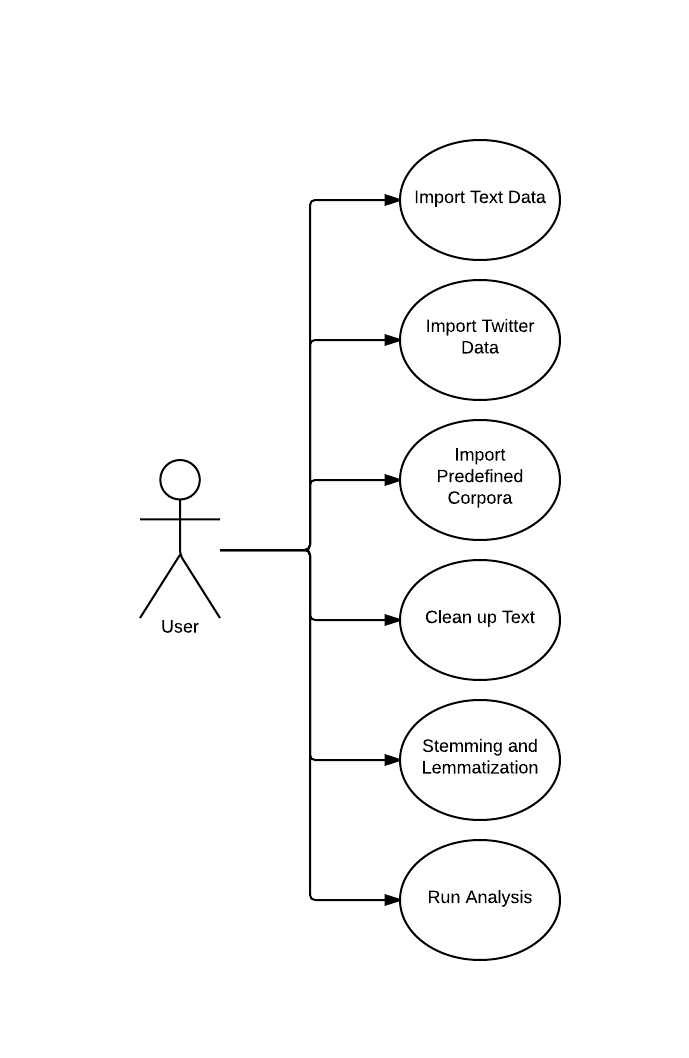
# Project Overview

* 1. Project Background
  2. Project Features
  3. User Classes

The primary target users of Linguine are students studying linguistics and natural language processing. This can include users who are experienced with NLP and computing/programming, users who are experienced with NLP but not programming, users who are experienced with programming but not NLP, and users who are experienced with neither NLP nor programming. Because Linguine is targeted at an educational environment and a wide range of user experience levels, usability is a key concern.

* 1. User Documentation
  2. Licensing

# Features and Use Cases



## 3.1 Import Text Data

### Description:

A user will import data by directly uploading a text file from their computer.

### Primary Flow:

1. The user shall indicate to the system their desire to import files.
2. The system shall provide the user with an interface to select one or more files or a single directory to import.
3. The user shall indicate one or more files or a directory.
4. The system shall upload the selected file(s) or directory and import them. The system will interpret the files as plaintext unless their extension is HTML or XML.

## 3.2 Import Twitter Data

### Description:

A user will import data from Twitter.

### Primary Flow:

1. The user shall indicate to the system their desire to import data from Twitter.
2. The system shall provide the user with an interface allowing the user to choose between importing general Twitter data or importing data related to their Twitter account.
3. The system shall import the selected corpus of tweets.
   1. If the user selected general Twitter data, the system shall query the Twitter API for tweets and return the results.
   2. If the user selected data from their Twitter account, the system shall prompt the user to log in with Twitter and return the results
      1. The system shall prompt the user for an authorization with Twitter
      2. If the user authorizes the action, the system will query Twitter’s API for tweets relevant to the user and return the resutls.

## 3.3 Import from Predefined Corpora

### Description:

A user will import data from a corpus that is defined by and included with the system.

### Primary Flow:

1. The user shall indicate to the system their desire to import a predefined corpus.
2. The system shall provide the user with an interface with which to select a corpus or individual texts. The included data shall be significant texts from the English language canon and a sample of journalistic articles.
3. The user shall select either a predefined corpus or an individual text that they wish to import.
4. The system shall import the selected corpus or text.

## 3.4 Export Session Status

### Description:

A user will export the entire contents of their session to a downloadable file.

### Primary Flow:

1. The user shall indicate to the system their desire to export the current session.
2. The system shall generate a file which represents the current status of the session, including all imported texts.
3. The user shall download the generated file.

## 3.5 Import Session Status

### Description:

A user will import a session that was previously exported to a file.

### Primary Flow:

1. The user shall indicate to the system their desire to import a session status file.
2. The system shall provide the user with an interface to select a single file.
3. The user shall select a single file (previously exported via use case 4).
4. The system shall restore the user’s session based on the information in the file.

## 3.6 Clean up of text

### Description:

A user will clean up imported data. “Clean up” is considered to include casing, whitespace, and punctuation.

### Prerequisites:

The user must have one or more texts imported.

### Primary Flow:

1. The system shall provide the user with an interface showing a collection of clean-up task options for the user to select from
2. The user shall select as many options as they wish to have the system perform during the cleaning process.
3. The user shall indicate to the system that they have selected the options they would like the system to use.
4. The system shall perform the selected processing tasks on the text and store the resultant text for other processing and analysis work.

## 3.7 Stemming and Lemmatization

### Description:

A user will stem or lemmatize one or more imported texts.

### Prerequisites:

The user must have one or more texts imported.

### Primary Flow:

1. The user will indicate to the system one or more imported texts (by selecting them).
2. The user will indicate to the system that they wish to stem or lemmatize data.
3. The system shall provide the user with an interface to select either stemming or lemmatization.
4. The user shall select either stemming or lemmatization.
5. The system shall either stem or lemmatize the text(s) indicated by the user in step 1.

## 3.8 Run Analysis

### Description:

The user shall run one or more analyses on one or more imported texts.

### Prerequisites:

One or more texts shall be imported into the system.

### Primary Flow:

1. The user shall indicate one or more imported texts by selecting them.
2. The user shall indicate to the system that they wish to run an analysis.
3. The system shall provide the user with a list of analyses that may be run. Possible analyses shall include sentence parsing, topic modelling, part-of-speech tagging, named entity recognition, and coreference identification.
4. The user shall select one or more analyses that they wish to run.
5. The system shall determine if any additional analyses need to be run as prerequisites to those selected by the user.
6. The system shall provide an interface to the user allowing them to select how they would like to visualize the output of their analysis.
7. The user shall indicate one or more output formats / visualizations.
8. The system shall run the requested analyses and any prerequisite analyses.
9. The system shall display the requested output and visualizations to the user.

# Nonfunctional Requirements

* 1. Usability

The application shall be easily learnable. A novice user shall be able to rapidly complete a basic use case of importing data and running a simple analysis without needing to reference documentation. The application shall also be easily learnable for users who are experienced with NLP analysis.

* 1. Testability

The application shall have a high percentage of unit test coverage. A more specific target percentage will be identified in the future. In addition to in-code testing, a test plan will be developed which will include usability testing with a sample of end users.

* 1. Web Architecture

The application shall utilize a web architecture. The primary language of implementation shall be Python.

# Appendix

# Functional Requirements and Use Cases

# Glossary

Corpus: A collection of associated texts.

Import: Add to the current session’s workbench.