Vorgehensweise:

Datenaufbereitung (Pre-processing Final):

* Händisch: Ändern der Textdatei in csv-Format
* Einlesen des Dokuments
* Entfernung von Satzzeichen und Tabulatorzeichen
* Entfernung der Stopwords (nltk.corpus)
* Synonyme finden mit Wordnet (nltk.corpus)
* Wortstammreduktion (nltk.stem.snowball SnowballStemmer)
* Wiederholung dieser Schritte mit allen Dokumenten

Klassifizierung (Classifying):

* Laden der aufbereiteten Trainingsdaten
* Zu einem Dataframe konkatenieren
* Vektorisieren mit TFIDF
* Test und Validation Sets (0,2) erstellen
* Accrurency Messung auf diesen mit
  + LogisticRegression 0,85
  + KNeighborsClassifier 0,75
  + DecisionTreeClassifier 0,69
  + GaussianNB 0,61
  + SVM 0,51
* Auswählen der besten Methode, hier LogisticRegression
* Auskommentieren der Methoden
* Laden aller aufbereiteten Daten
* Dataframe erstellen und TFIDF
* LR mit Trainingsdaten fitten
* Ungelabelte Daten voraussagen