Proces

We zijn begonnen met het inlezen van de data in een panda dataframe.

Hier hebben we eerst verschillende kolommen uit gehaald om te kijken hoe python deze zou clusteren.

We willen kmeans gebruiken omdat dit een snelle manier van clusteren is.

Grote outlier gevonden in de dataset. Warm lentil salad with Serrano chicken en rocket heeft 2012 carbs wat ons erg onwaarschijnlijk lijkt. Daarom hebben we besloten de andere outliers ook uit de dataset te gooien.

4/4/2018

Alle missende waardes nan eruit gegooid.

Proberen dubbele waardes eruit te filteren door de functie data.duplicate.

De nutritions uit de dataset verwijderen

17/4/2018

Stopwoorden:

Stopwoorden wegfilteren met behulp van nltk stopwords package. Dit hebben we gedaan voor 4 categorieën: title description ingredients en methods. De output hiervan waren losse woorden in een list per rij. Deze zijn later weer samengevoegd tot een regel per rij en terug in een df gestopt. Dit is makkelijk voor het gebruik van een bag of words. Tot slot is de term frequency inversed document frequency (tfidf).

De categorie skill level is omgezet tot een categorie met integers. Dit zijn 3 levels, 0: easy, 1: more effort, 2: challenge.

In het maken van de cluster\_data met een dictionary, werd aan elke rij alle waardes toegevoegd omdat het een grote list was. Met de volgende code:

cluster\_data = pd.DataFrame({'title':tfidf\_title, 'description':tfidf\_descr, 'ingredients':tfidf\_ingr, 'method':tfidf\_meth, 'skill\_level':f\_enc[0]})

cluster\_data\_array = cluster\_data.values

Dit is met een for-loop opgelost door voor elke rij een list aan te maken waardoor de data gesplitst wordt en er bij elke index de juiste waarde wordt gekoppeld.