TÉCNICAS AVANZADAS DE ANÁLISIS DE DATOS

Recommender systems

Lubomira Smolarova

Introduction

In this scenario, the client has a question or search query, and they have a set of documents available. By utilizing KNN, the algorithm can calculate the similarity between the query and the documents based on their vector representations, and identify the k nearest neighbors (documents) that are most similar to the query.

The k-nearest neighbors (KNN) algorithm can be used for ranking based on TF (Term Frequency) and TF-IDF (Term Frequency-Inverse Document Frequency), which are commonly used methods for representing textual data. TF measures the frequency of terms in a document, while TF-IDF combines TF with IDF (Inverse Document Frequency) to assess the importance of terms across a collection of documents.

Another representation called Bag-of-Words (BoW) is closely related to the Term Frequency (TF) representation. Both BoW and TF aim to capture the frequency of terms in a document. TF represents the raw count of each term in a document, whereas BoW takes into account the occurrence of terms across all documents in the collection. BoW represents a document as a vector of term frequencies, where each entry in the vector corresponds to the frequency of a specific term in the document.

Analysis of feature extraction techniques

Experiments on Document.text from the assignment and K= 2. Analysis of TF-IDF, TF and BoW differences as feature extraction techniques. First, I set the random seed for reproducibility. Following are the queries and corresponding highest 2 ranking documents. There are 7 documents in total. Experiments are done on 5 queries: "Fruity and balanced wine", "Crisp and refreshing white wine", "Smooth and medium-bodied red wine", "Aromatic and vibrant wine" and "Dry and elegant wine".

Ranking based on TF-IDF

Query: Fruity and balanced wine
Rank 1: This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016.
Rank 2: Here's a bright, informal red that opens with aromas of candied berry, white pepper and savory herb that carry over to the palate. It's balanced with fresh acidity and soft tannins.

Query: Crisp and refreshing white wine
Rank 1: Tart and snappy, the flavors of lime flesh and rind dominate. Some green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented.
Rank 2: Here's a bright, informal red that opens with aromas of candied berry, white pepper and savory herb that carry over to the palate. It's balanced with fresh acidity and soft tannins.

Query: Smooth and medium-bodied red wine
Rank 1: This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016.
Rank 2: Blackberry and raspberry aromas show a typical Navarran whiff of green herbs and, in this case, horseradish. In the mouth, this is fairly full bodied, with tomatoey acidity. Spicy, herbal flavors complement dark plum fruit, while the finish is fresh but grabby.

Query: Aromatic and vibrant wine
Rank 1: Tart and snappy, the flavors of lime flesh and rind dominate. Some green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented.
Rank 2: This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016.

Query: Dry and elegant wine
Rank 1: Tart and snappy, the flavors of lime flesh and rind dominate. Some green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented.
Rank 2: This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016.

Ranking based on TF

Query: Fruity and balanced wine
Rank 1: Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.
Rank 2: Here's a bright, informal red that opens with aromas of candied berry, white pepper and savory herb that carry over to the palate. It's balanced with fresh acidity and soft tannins.

Query: Crisp and refreshing white wine
Rank 1: Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.
Rank 2: Here's a bright, informal red that opens with aromas of candied berry, white pepper and savory herb that carry over to the palate. It's balanced with fresh acidity and soft tannins.

Query: Smooth and medium-bodied red wine
Rank 1: Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.
Rank 2: Here's a bright, informal red that opens with aromas of candied berry, white pepper and savory herb that carry over to the palate. It's balanced with fresh acidity and soft tannins.

Query: Aromatic and vibrant wine
Rank 1: Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.
Rank 2: Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semidry finish.

Query: Dry and elegant wine
Rank 1: Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.
Rank 2: Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semidry finish.

Ranking based on Bag-of-Word

Query: Fruity and balanced wine
Rank 1: Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.
Rank 2: Here's a bright, informal red that opens with aromas of candied berry, white pepper and savory herb that carry over to the palate. It's balanced with fresh acidity and soft tannins.

Query: Crisp and refreshing white wine
Rank 1: Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.
Rank 2: Here's a bright, informal red that opens with aromas of candied berry, white pepper and savory herb that carry over to the palate. It's balanced with fresh acidity and soft tannins.

Query: Smooth and medium-bodied red wine
Rank 1: Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.
Rank 2: Here's a bright, informal red that opens with aromas of candied berry, white pepper and savory herb that carry over to the palate. It's balanced with fresh acidity and soft tannins.

Query: Aromatic and vibrant wine
Rank 1: Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.

Rank 2: Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semidry finish.

Query: Dry and elegant wine
Rank 1: Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.
Rank 2: Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semi dry finish.

Table of first four ranks for each query

	"Crisp and refreshing white wine"	"Crisp and refreshing white wine"	"Smooth and medium-bodied red wine"	"Aromatic and vibrant wine"	"Dry and elegant wine"
TF-IDF	2 7 3 5	3 7 2 5	2 6 7 3	3 2 5 1	3 2 5 1
TF	1 7 4 3	1 7 3 4	1 7 4 3	1 4 7 3	1 4 7 3
BoW	1 7 4 3	1 7 3 4	1 7 4 3	1 4 7 3	1 4 7 3

Conclusion

Reading TF-IDF results, its recommendations are pretty good. TF on the other hand always recommends the same 2 documents. BoW for the first three queries recommends the same documents and for the last two recommends a one different document.

During analysis of first 4 ranks of recommendations TF and BoW were exactly the same. Document number 3 was in the top four for each query and each FE. Document number 7 appeared 13/15 times. Never as the first one recommended.