Exercises: Text Classification and Clustering

September 27, 2016

Task 1: Preprocessing

Preprocess the documents in Table 1 and construct the corresponding document-term matrix. Table 2 contains the stopwords list to be used. Apply *suffix-s* stemming.

docID	content
1	The King's Speech
2	The Lord of the Rings: The Return of the King
3	Street Kings
4	The Scorpion King
5	The Lion King

Table 1: Document collection consisting of movie titles.

a	as	by	into	not	such	then	this	with
an	at	for	is	of	that	there	to	
and	be	if	it	on	the	these	was	
are	but	in	no	or	their	they	will	

Table 2: Standard English stopwords list.

Solution

DocID	king	speech	lord	ring	return	street	scorpion	lion
1	1	1						
2	1		1	1	1			
3	1					1		
4	1						1	
5	1							1

Table 3: Document-term matrix.

Task 2: K-Nearest Neighbor Classification

Find $K = 2$ nearest neighbors of document 1 using the Jaccard coefficient and the cosine similarity.

$$cos(d_1, d_2) = \frac{\sum_t n(t, d_1) n(t, d_2)}{\sqrt{\sum_t n(t, d_1)^2 \sum_t n(t, d_2)^2}} \tag{1}$$

Solution

Documents	Jaccard	Cosine
1 vs. 2	0.2	0.35
1 vs. 3	0.33	0.5
1 vs. 4	0.33	0.5
1 vs. 5	0.33	0.5

Table 4: Similarity between document 1 and documents 2–5.

Task 3: Naive Bayes Classification

Assuming a binary classification task, where the documents are labeled as shown in Table 5, how would the document d ="The Mummy King Returns" be classified using a Naive Bayes classifier? Use Laplace smoothing for computing term probabilities.

docID	**target (Oscar?)**
1	Yes
2	Yes
3	No
4	No
5	Yes

Table 5: Target labels for the document collection

		$n(t,c)$								
Class	N_c	king	speech	lord	ring	return	street	scorpion	lion	SUM
Oscar=Yes	3	3	1	1	1	1			1	8
Oscar=No	2	2					1	1		4

Table 6: Term counts.

$$P(c|d) \propto P(c) \prod_{t \in d} P(t|c)^{n(t,d)} \tag{2}$$

where

$$P(c) = \frac{N_c}{N} \qquad P(t|c) = \frac{n(t,c)+1}{\sum_t n(t,c) + |V|} \tag{3}$$

Solution

$P(d|Yes) = P(Yes) \cdot P(mummy|Yes) \cdot P(king|Yes) \cdot P(return|Yes) = \frac{3}{5} \cdot \frac{0+1}{8+8} \cdot \frac{3+1}{8+8} \cdot \frac{1+1}{8+8} = 0.00117$

$P(d|No) = P(No) \cdot P(mummy|No) \cdot P(king|No) \cdot P(return|No) = \frac{2}{5} \cdot \frac{0+1}{4+8} \cdot \frac{2+1}{4+8} \cdot \frac{0+1}{4+8} = 0.00069$

$P(d|Yes) > P(d|No)$ therefore it will be classified as "Yes."

Task 4: K-Nearest Neighbors Clustering

Perform the first 2 iterations of K-Nearest Neighbors clustering with $K = 2$ for the documents in Table 1. Use Jaccard similarity. Take documents 1 and 5 as the initial centroids.

Solution

Round numbers to the first digit.
The underline indicates which centroid the point was assigned to. In case of draws (equal distance from both cluster centroids) we made an arbitrary choice. The cluster centroids in the next iteration are updated based on the points assigned to that cluster.
Mind that Jaccard similarity is for binary attributes: either a term is present in a document or not. When computing similarities in Iteration 2, we consider values $\geq \frac{1}{2}$ as the term being present ("1") and values $< \frac{1}{2}$ as the term being absent ("0"). The 'binarized' vectors are shown in the second row in the Iteration 2 heading of the table.

DocID	Term vector	Iteration 1 Distance to centroid of Cluster 1 (1,1,0,0,0,0,0,0)	Iteration 1 Distance to centroid of Cluster 2 (1,0,0,0,0,0,0,1)	Iteration 2 Distance to centroid of Cluster 1 $(1,\frac{1}{3},\frac{1}{3},\frac{1}{3},\frac{1}{3},\frac{1}{3},0,0)$ (1,0,0,0,0,0,0,0)	Iteration 2 Distance to centroid of Cluster 2 $(1,0,0,0,0,0,\frac{1}{2},\frac{1}{2})$ (1,0,0,0,0,0,1,1)
1	(1,1,0,0,0,0,0,0)	$\underline{1}$	$\frac{1}{3}$	$\underline{\frac{1}{2}}$	$\frac{1}{4}$
2	(1,0,1,1,1,0,0,0)	$\underline{\frac{1}{5}}$	$\frac{1}{5}$	$\underline{\frac{1}{4}}$	$\frac{1}{6}$
3	(1,0,0,0,0,1,0,0)	$\underline{\frac{1}{3}}$	$\frac{1}{3}$	$\underline{\frac{1}{2}}$	$\frac{1}{4}$
4	(1,0,0,0,0,0,1,0)	$\frac{1}{3}$	$\underline{\frac{1}{3}}$	$\frac{1}{2}$	$\underline{\frac{2}{3}}$
5	(1,0,0,0,0,0,0,1)	$\frac{1}{3}$	$\underline{1}$	$\frac{1}{2}$	$\underline{\frac{2}{3}}$

Table 7: K-Nearest Neighbors Clustering

Task 5: Hierarchical Agglomerative Clustering

Perform the first step of the Hierarchical Agglomerative Clustering method for the documents in Table 1. Use cosine similarity and the group average cluster proximity.

Solution

Init: Compute the proximity matrix
Each document corresponds to a (singleton) cluster.

	doc 1	**doc 2**	**doc 3**	**doc 4**	**doc 5**
doc 1	0	$\frac{1}{\sqrt{8}}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
doc 2		0	$\frac{1}{\sqrt{8}}$	$\frac{1}{\sqrt{8}}$	$\frac{1}{\sqrt{8}}$
doc 3			0	$\frac{1}{2}$	$\frac{1}{2}$
doc 4				0	$\frac{1}{2}$
doc 5					0

Table 8: Initial proximity matrix.

First iteration: Merge the two most similar clusters and update the proximity matrix

The merged frequency vector for doc 1,3 becomes (2,1,0,0,0,1,0,0).

	doc 1,3	**doc 2**	**doc 4**	**doc 5**
doc 1,3	0	$\frac{2}{\sqrt{24}}$	$\frac{2}{\sqrt{12}}$	$\frac{2}{\sqrt{12}}$
doc 2		0	$\frac{1}{\sqrt{8}}$	$\frac{1}{\sqrt{8}}$
doc 4			0	$\frac{1}{2}$
doc 5				0

Table 9: Proximity matrix after the first iteration.