Выделение признаков из текста. Начало

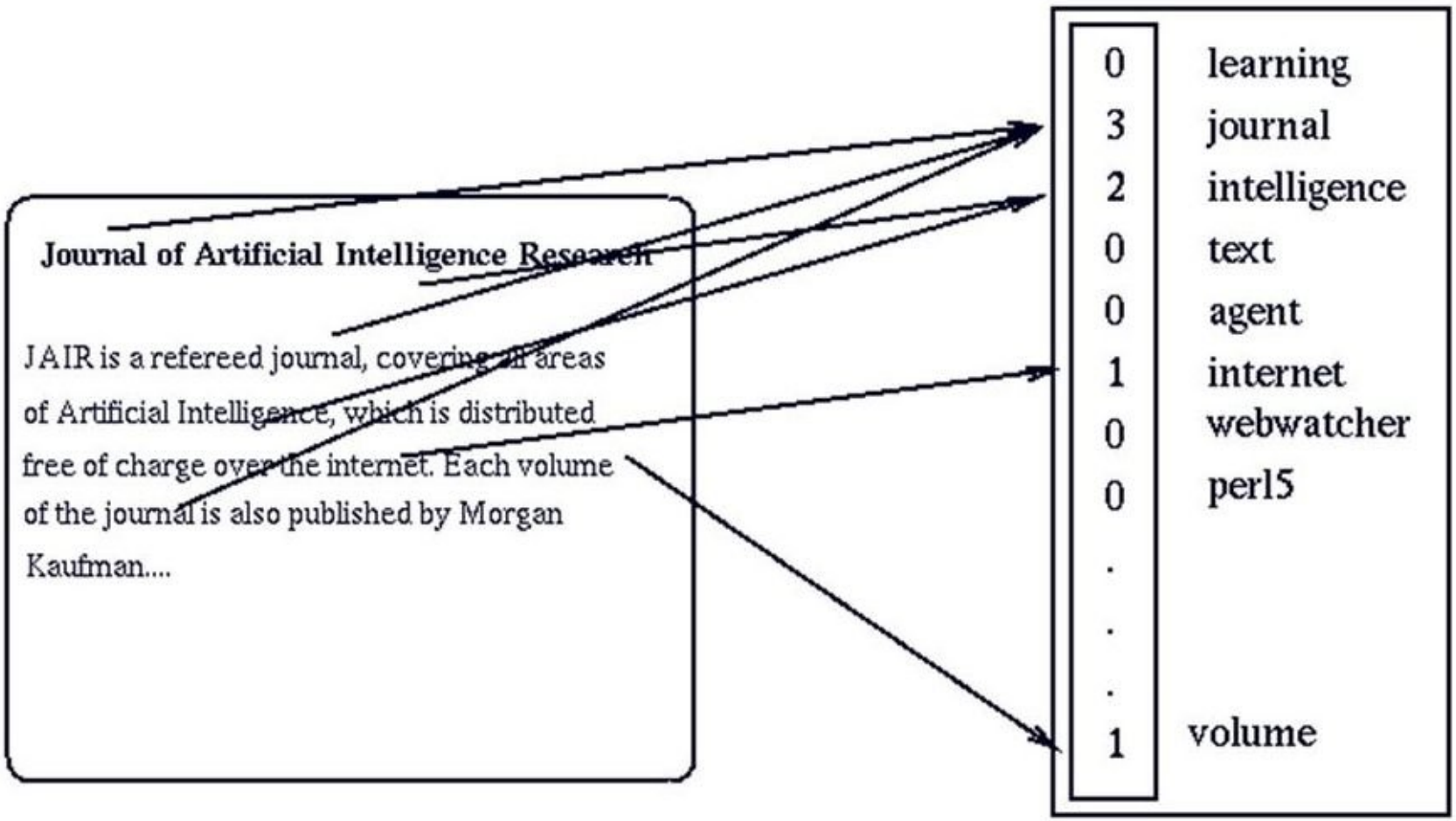
Выступающий: Арина Агеева

ФКН ВШЭ

Москва, 2017

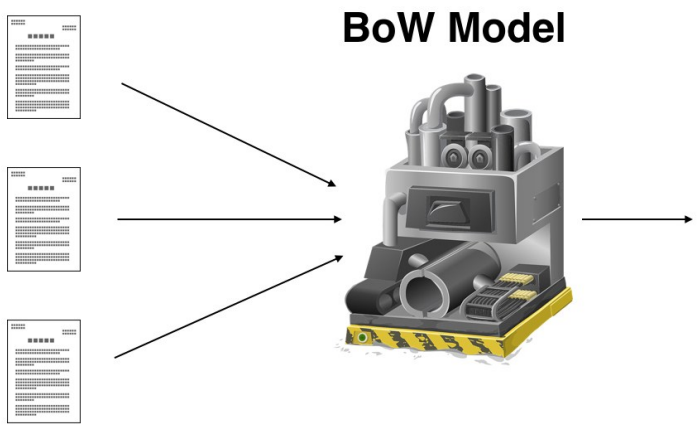
Содержание

Мешок слов (BOW)



Пример

1. I love dogs.
2. I hate dogs and knitting.
3. Kinitting is my hobby and my passion



	I	love	dogs	hate	and	knittin	is	my	hobby	passi
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

Содержание

Частота слова (TF – term frequency)

Пусть $w \in V$ – словарь,
$d \in D$ – корпус документов
Тогда

$$tf(w, d) = \frac{n_{wd}}{n_d}$$

где n_{wd} – число вхождений слова w в документ d,
$n_d = \sum_{w \in V} n_{wd}$ – общее число слов в документе

Оценивается важность некоторого слова в пределах одного документа

Обратная частота документа (IDF – inverse document frequency)

Пусть $w \in V$ – словарь, D – корпус документов
Тогда

$$idf(w, D) = \log \frac{|D|}{|\{d \in D | w \in d\}|}$$

где
$|D|$ – число документов в коллекции,
$|\{d \in D | w \in d\}|$ – число документов из коллекции D, в которых встречается w

Учёт IDF уменьшает вес широкоупотребительных слов

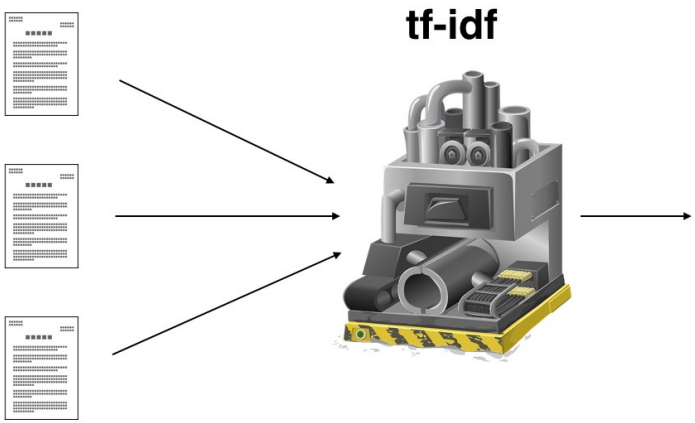
Мера TF-IDF

$$\text{tf-idf}(w, d, D) = tf(w, d) idf(w, D)$$

Большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах

Пример

1. I love dogs.
2. I hate dogs and knitting.
3. Kinitting is my hobby and my passion



	I	love	dogs	hate	and	knittin	is	my	hobby	passi
Doc 1	0.18	0.48	0.18							
Doc 2	0.18		0.18	0.48	0.18	0.18				
Doc 3					0.18	0.18	0.48	0.95	0.48	0.48

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	0.18	**0.48**	0.18							
Doc 2	0.18		0.18	**0.48**	0.18	0.18				
Doc 3					0.18	0.18	**0.48**	**0.95**	**0.48**	**0.48**

Спасибо за внимание!