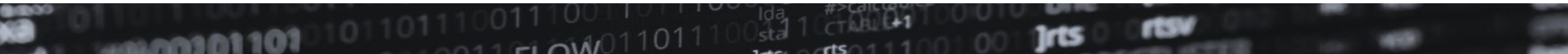
Распознавание языка

11.02.17
семинар



ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Методы распознавания языка

1. Методы
   - Символьные n-граммы
   - Частотные слова
2. Классификатор:
   - признаки

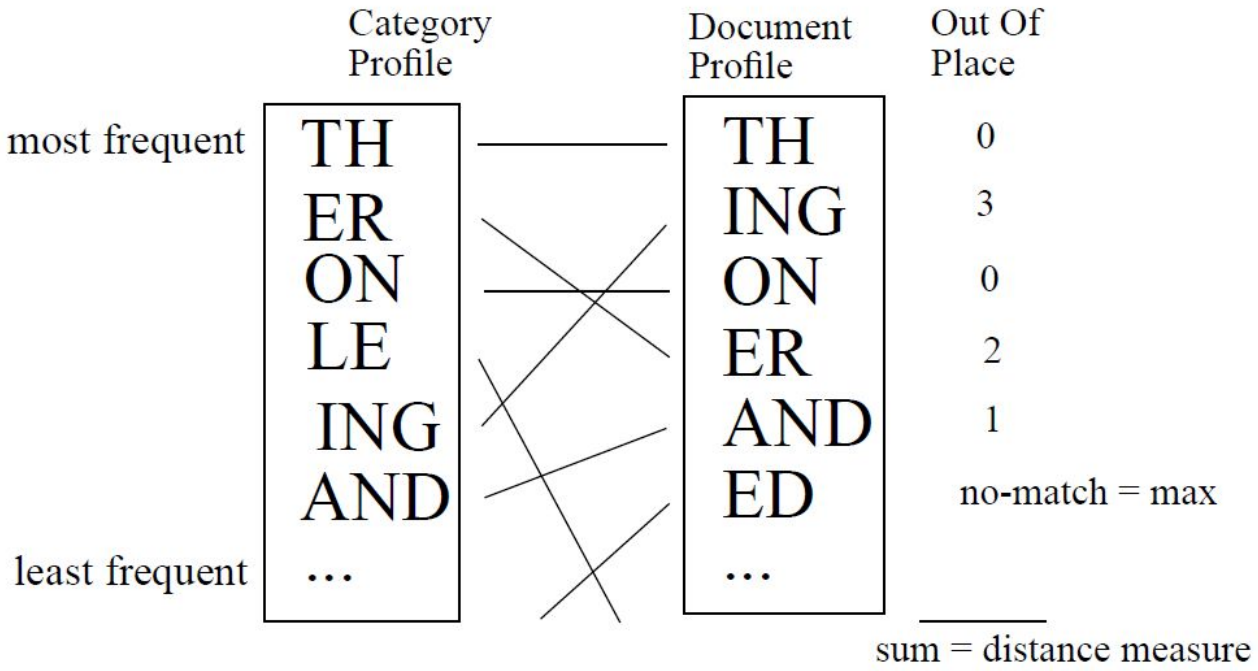
Методы распознавания языка: n-граммы

- Символьные n-граммы
- Допустим, у меня есть текст text
  - биграммы: _T, TE, EX, XT, T_
  - триграммы: _TE, TEX, EXT, XT_, T_ _
  - квадрограммы: _TEX, TEXT, EXT_, XT_ _, T_ _ _

Методы распознавания языка: n-граммы

- Препроцессинг: удаление пунктуации и служебных знаков, регуляризация пробелов, токенизация
- Инвертированный по частоте порядок n-грамм
- Нормализация
- Ципф
- Можно взять 300 самых частотных n-грамм, остальное выкинуть

Методы распознавания языка: n-граммы



http://blog.alejandronolla.com/2013/05/20/n-gram-based-text-categorization-categorizing-text-with-python/

Методы распознавания языка: n-граммы

- Можно считать вероятности n-грамм (логарифмы)

	German	English	French	Danish	Swedish
0x5468 'Th'	-8.67	-6.37	-9.60	-9.40	-10.00
0x6865 'he'	-4.96	-4.08	-6.51	-6.03	-6.23
0x6520 'e '	-3.94	-3.56	-3.24	-3.71	-4.74
0x2063 ' c'	-10.10	-5.00	-4.69	-7.66	-7.78
0x6361 'ca'	-9.46	-5.89	-6.28	-8.27	-8.47
0x6174 'at'	-5.63	-4.80	-5.37	-5.04	-4.81
total:	-42.79	-29.73	-35.72	-40.13	-42.05

The cat

Для английского вероятность предложения - (-6.37) + (-4.08) + (-3.56) + (-5.00) + (-5.89) + (-4.80) = **-29.73**, самая высокая вероятность из пяти языков

http://www.practicalcryptography.com/miscellaneous/machine-learning/tutorial-automatic-language-identification-ngram-b/

Методы распознавания языка: словарный метод

- Корпуса по 1 млн. слов
- Строим частотный список

Пример - Ein:

- немецкий - 345032/46387276 = 0.0074 (около 0.74 % от всех немецких слов)
- английский – 9 / 57699127 words

http://practicalcryptography.com/miscellaneous/machine-learning/tutorial-automatic-language-identification-word-ba/

Методы распознавания языка: словарный метод

	German	English	French	Danish	Swedish
THE	-9.21	-2.76	-9.01	-8.54	-8.87
CAT	-13.7	-10.83	-13.27	-12.81	-12.13
PLAYS	-15.34	-9.41	-16.04	-15.25	-15.56
total:	-38.25	-23.01	-38.33	-36.61	-36.56

The cat plays

Английский: (-2.76) + (-10.83) + (-9.41) = -23.01

http://practicalcryptography.com/miscellaneous/machine-learning/tutorial-automatic-language-identification-word-ba/

Методы распознавания языка: словарный метод

- В тестовой выборке количество предложений из каждого языка: 10000
- 50,000 тестовых предложений, точность 99.84% (49924 из 50000)
- Guessed label

		de	en	fr	da	sw
Actual Labels	de	9995	3	1	1	0
	en	1	9998	1	0	0
	fr	2	7	9990	0	1
	da	2	1	0	9994	3
	sw	1	2	2	48	9947

http://practicalcryptography.com/miscellaneous/machine-learning/tutorial-automatic-language-identification-word-ba/

ШКОЛА ЛИНГВИСТИКИ НИУ ВШЭ

Автоматическая классификация

- Объекты (множество прецедентов)
  - Текст на языке L
- Признаки
  - N-граммы / bag of words
  - N-граммы - *сим, имв, мво, вол, оло, лов*
- Отклик (целевые классы)
  - Язык X / не язык X
  - Множество языков
- Классификатор: наивный Байес
- https://bitbucket.org/Smyek/language-identification/