## ДЗ №1 Специфичность лексики

Стырина Соня

2 общеупотребительных слова:

day world

2 специфических слова для первого подкорпуса:

artist exhibition

**Specific Corpus**: e-flux (International art English)

1) тематика: собрание новостей об искусстве и выставках

2) источник текстов: [e-flux.com](http://e-flux.com)

3) объем корпуса: 5,036,119 словоупотреблений

4) токенизирован с помощью TreeTagger и Penn Treebank tagset

**Reference Corpus**: Brexit corpus without retweets (English)

1) тематика: новостные статьи, твиты и посты в блогах, обсуждающие Brexit

2) источник текстов: новостные сайты (The Guardian, BBC, the Daily Mail, the Telegraph, etc.), твиттер, блоги

3) объем корпуса: 4,789,571 словоупотреблений

4) токенизирован с помощью TreeTagger и Penn Treebank tagset & Sketch Engine modifications

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| wi | Тип (тематическое vs. общеупотребительное) | CountSpecC  (частота в Specific Corpora) | CountRefC (частота в контрастном корпусе  (Reference Corpora)) | LogLikelihood | Ранг | Вес 2 (wiredness) | Ранг |
| day | общеупотребительное | 2377 | 5747 | 1615.35 | 3 | 0.39 | 4 |
| world | общеупотребительное | 7622 | 6122 | 97.40 | 4 | 1.18 | 3 |
| exhibition | тематическое | 32542 | 25 | 43127.18 | 2 | 1237.96 | 1 |
| artist | тематическое | 34087 | 161 | 43749.30 | 1 | 201.36 | 2 |

**Второй метод**: Wiredness = (wi) = frs(wi)/frr(wi) = (Ws / Ts) / (Wr / Tr)

frs(wi) – относительная частота слова в коллекции текстов определенной тематической области

frr(wi) – относительная частота слова в контрастной коллекции (reference corpus)

Ws – абсолютная частота wi в тематической коллекции

Ts – количество слов в тематической коллекции

Wr – абсолютная частота слова в контрастной коллекции (reference corpus)

Tr – количество слов в контрастной коллекции (reference corpus)

(источник: лекция 1)

Комментарии:

Оба метода поместили 2 тематических слова на высокие позиции в ранжированном списке, но 1 и 2 ранг были распределены по-разному:

LogLikelihood → ранг 1: artist

Wiredness → ранг 1: exhibition