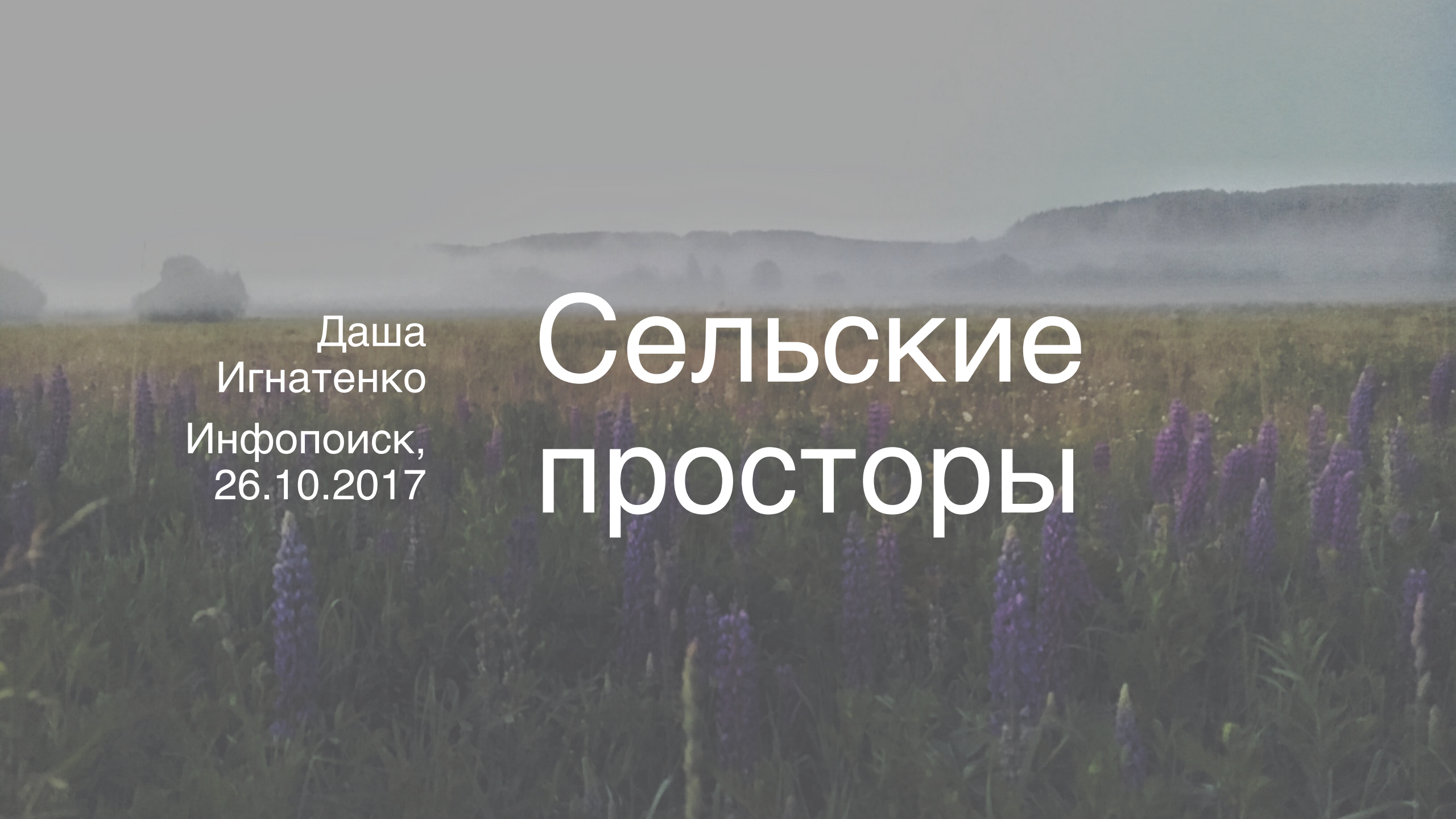


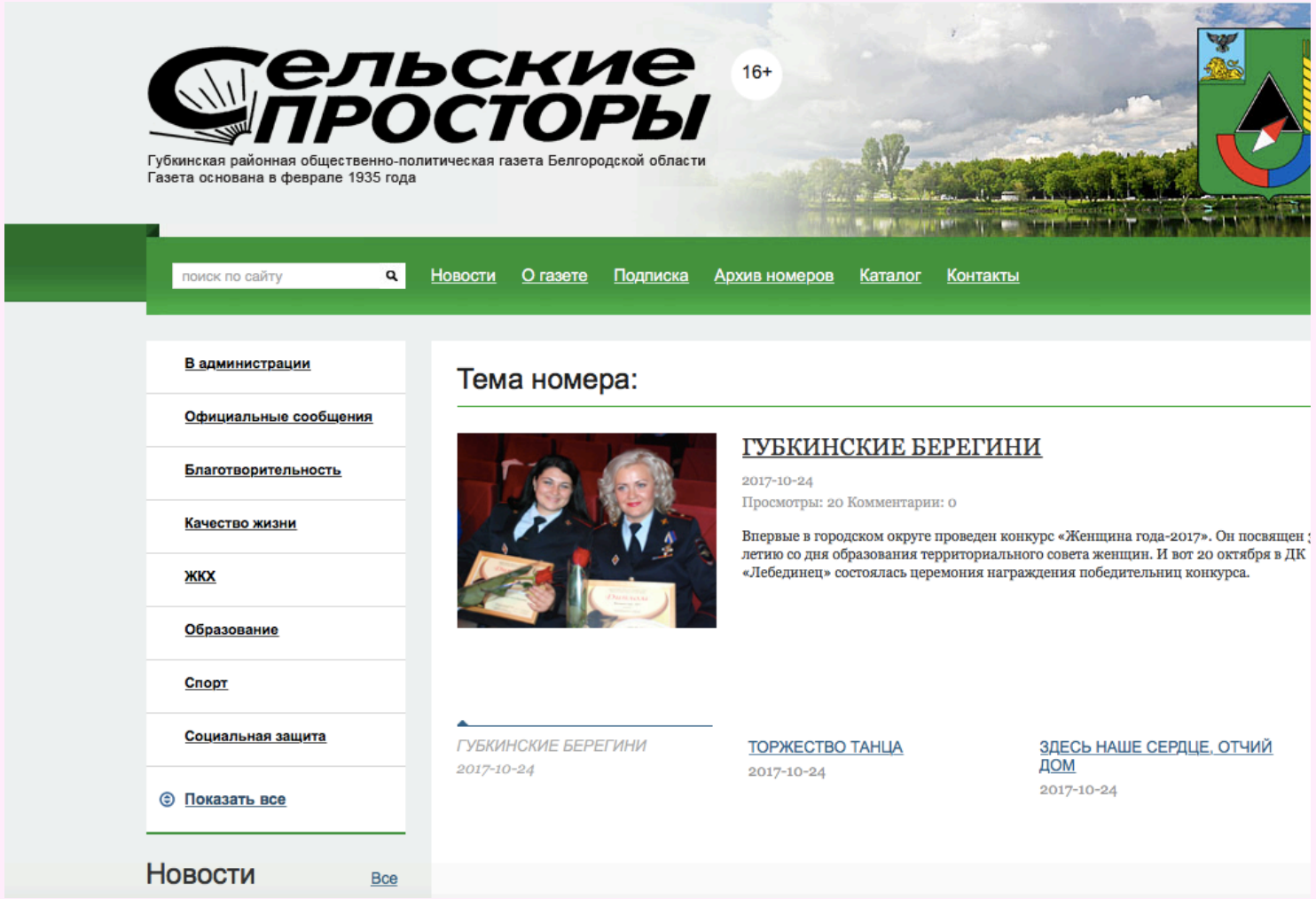
Сельские просторы

Даша Игнатенко

Инфопоиск, 26.10.2017

Ресурс

- http://sprostor.ru



Ресурс

- http://sprostor.ru
- Парсинг через раздел с архивом
- Краулер:
  - Собирает по всем страницам ссылки на все выпуски
  - Проходит по выпускам и собирает ссылки на статьи
  - Выкачивает статьи, записывая метаданные

Корпус

- 1710 документов
- 30503 лемм
  - - стоп-слова
  - Лемматизатор из Mystem
- 804062 токенов

Корпус

- 1710 документов
- Templates
  - Index.html
- App.py

Обратные индексы

In [3]:
```
def indices():
    d = {}
    for root, dirs, files in os.walk('.'):
        for file in files:
            if file.endswith('.txt'):
                f = open(file, 'r', encoding='utf-8')
                lemmas = m.lemmatize(f.read())
               # for i in set(lemmas):
                for i in lemmas:
                    if i not in stopw:
                        if i not in d:
                            d[i] = [file]
                        else:
                            d[i].append(file)
                f.close()
    return(d)
```

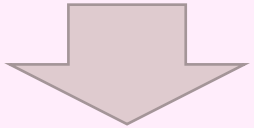
In [4]:
```
d = indices()
```

In [8]:
```
d
```

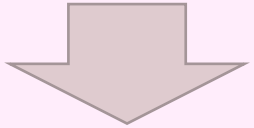
Out[8]:
```
{'дарина': ['1587.txt', '928.txt', '937.txt'],
 'кубанский': ['234.txt'],
 'подпрограмма': ['1029.txt',
  '112.txt',
  '1366.txt',
  '1657.txt',
  '1657.txt',
  '1657.txt',
  '362.txt',
  '487.txt',
  '487.txt',
```

Ранжирование – топ 10

- Функция, которая возвращает Okapi score для 1 слова во всех документах (словарь)



- Функция, которая обрабатывает запросы любой длины и возвращает 10 самых релевантных



- Ещё одна функция, которая переводит выдачу в формат (url, заголовок статьи, okapi score)

Ранжирование – топ 10

```
In [28]: search('кот, собака?')

Out[28]: [(' http://sprostor.ru/?module=articles&action=view&id=1690',
           0.218662847687655 98),
          (' http://sprostor.ru/?module=articles&action=view&id=3419',
           0.123768289162391 76),
          (' http://sprostor.ru/?module=articles&action=view&id=3546',
           0.122341090239652 73),
          (' http://sprostor.ru/?module=articles&action=view&id=2121',
           0.096106432937204 06),
          (' http://sprostor.ru/?module=articles&action=view&id=2786',
           0.052483123859339786),
          (' http://sprostor.ru/?module=articles&action=view&id=2769',
           0.048956349262272304),
          (' http://sprostor.ru/?module=articles&action=view&id=2083',
           0.034945613334551 38),
          (' http://sprostor.ru/?module=articles&action=view&id=3669',
```

•°*”~~”*°• Дизайн•°*”~~”*°•

Your query: ложка

Search

ДЛЯ ЛЕНИВЫХ И НЕ ОЧЕНЬ 0.2645586050431903 4

ВАМ ПОНРАВИТСЯ! 0.21604068283104497

Хозяйке на заметку 0.17912189232300055

ВСЯ АПТЕКА – В ЛОЖКЕ МЕДА 0.11998628188792182

ПОПРОБУЙ АНКЛ-БЭНС ПО-ТРОИЦКИ 0.10878626813062159

ВОТ ТАК «БУРЖУЙ»! 0.050188885575346

ЛУЧШЕ КАШИ БЛЮДА НЕ НАЙТИ 0.044207343748308 2

ЭПИДЕМИЧЕСКИЙ СЕЗОН: ЗАБОЛЕВАЕМОСТЬ СНИЖАЕТСЯ 0.029615373241185645

ЛЮБИТ РЫЖИК ТЫКВУ 0.01502595055965891 3

УРОКИ ЗЕЛЕНОГО ЦВЕТА 0.010010189789619