**Министерство образования и науки Российской Федерации Федеральное государственное автономное образовательное учреждение высшего образования**

**Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики**

**Факультет «Инфокоммуникационных технологий» Кафедра «Интеллектуальных технологий в гуманитарной сфере»**

**Направление подготовки «09.04.02 ̶ Информационные системы и технологии» Магистерская программа «Информационные системы управления в гуманитарной сфере»**

# ОТЧЕТ ПО ДИСЦИПЛИНЕ «ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ WEB - ДАННЫХ»

**Тема «Анализ веб-портала о глубинном обучении (deeplearning)»**

**Выполнил:**

Студент группы К4240

Баранов Я.

# Проверил:

Хлопотов Максим Валерьевич

Санкт-Петербург 2018

# Данные о проекте

1. Тема проекта: – анализ содержимого информационного веб-портала о глубинном обучении или deeplearning.
2. Бизнес-требования:
   1. Цель проекта: проанализировать содержимое веб-портала с целью выделения наиболее релевантных топиков по теме проекта (инженерное образование) и составить общую картину о целесообразности использования данного портала в качестве информационного ресурса.
   2. Значения (кто будет использовать проект): сотрудники образовательных учреждений (ВУЗы и школы), специалисты в области разработки и развития программ в области глубинного обучения.
   3. Ограничения (почему будут использовать этот проект): позволит более грамотно и удобно анализировать портал, что позволит минимизировать время, которое потребуется для изучения информации по теме.

По стандарту CRISP-DM включает следующие фазы:

* Понимание бизнеса (Business understanding).
* Понимание данных (Data understanding).
* Подготовка данных (Data preparation).
* Моделирование (Modeling).
* Оценка результатов (Evaluation).

# Понимание бизнеса

Анализ веб-порталов с помощью парсинга html страниц, позволит более понимать структуру содержимого сайта и минимизировать время на изучение содержащейся на нем информации.

# Понимание данных

В качестве данных используется набор html страниц с сайта, которые представлены в виде списка текстовых векторов для языка R.

# Подготовка данных

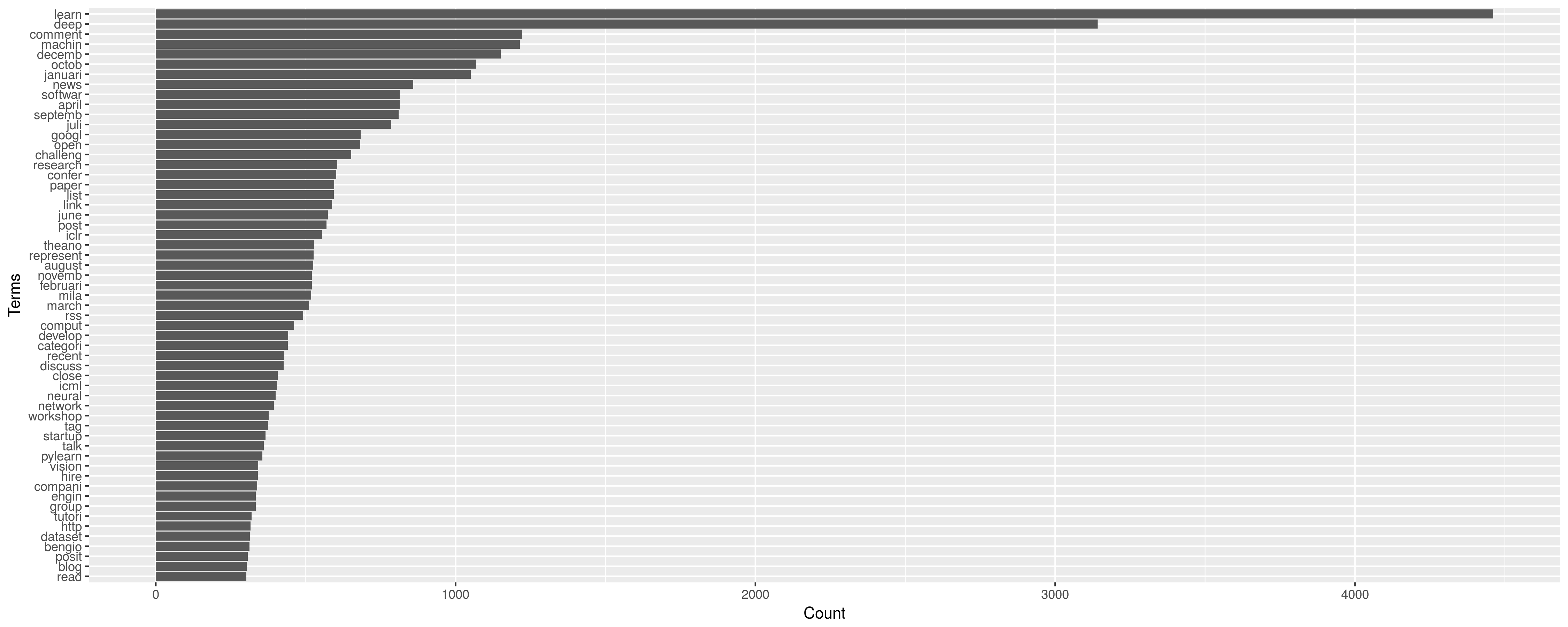
После скачивания, сырые данные их html страниц нужно обработать, чтобы смоделировать корректный корпус слов: очистить данные от тегов языка разметки, знаков переноса строки, знаков табуляции, убрать знаки пунктуации, удалить стоп-слова и удалить пробелы из корпуса слов

# Моделирование

Для моделирования тем или топиков (topics) был использован метод Латентного распределения Дирихле (Latent Dirichle Allocation) c помощью создания матрицы терминов или термов(as.DocumentTermMatrix) из корпуса текстов и запуск моделирования топиков с помощью функции LDA с указанием количества топиков до 5 как рекомендуемого создателями пакета.

# Оценка результатов

С помощью проведенного анализа сайта http://deeplearning.net/ посвященного изучению глубинного обучения в качестве результатов можно представить ряд визуализаций:

Рис.1 Распределение частот слов в текстовом содержимом сайта.

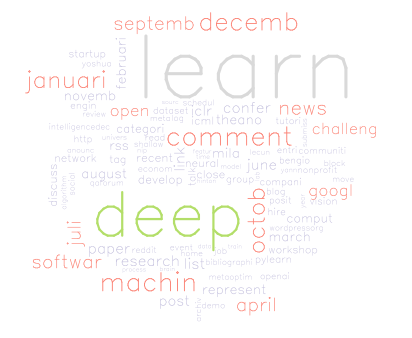


Рис. 2 Облако слов

Также можем увидеть самые часто встречающие слова в нашем корпусе текстов:

terms freq

april april 814

august august 525

blog blog 303

categori categori 440

challeng challeng 652

close close 407

comment comment 1222

compani compani 338

comput comput 461

confer confer 602

dataset dataset 314

decemb decemb 1150

deep deep 3141

develop develop 442

discuss discuss 427

engin engin 334

februari februari 520

googl googl 683

group group 333

hire hire 340

iclr iclr 554

icml icml 404

januari januari 1051

juli juli 785

june june 574

learn learn 4461

link link 588

list list 594

machin machin 1215

march march 511

mila mila 518

news news 859

novemb novemb 520

octob octob 1068

open open 682

paper paper 595

posit posit 307

post post 569

pylearn pylearn 356

read read 302

recent recent 429

represent represent 526

research research 605

rss rss 492

septemb septemb 810

softwar softwar 814

startup startup 366

tag tag 374

talk talk 360

theano theano 527

tutori tutori 319

vision vision 342

bengio bengio 313

http http 316

network network 394

neural neural 400

workshop workshop 376

И топики, которые мы получили в результате моделирования с помощью LDA:

Topic 1 Topic 2 Topic 3 Topic 4 Topic 5

[1,] "theano" "learn" "learn" "learn" "sourc"

[2,] "pylearn" "deep" "deep" "research" "theano"

[3,] "code" "neural" "machin" "machin" "compil"

[4,] "python" "network" "decemb" "deep" "memori"

[5,] "releas" "machin" "challeng" "comput" "set"

[6,] "document" "bengio" "octob" "announc" "optim"

[7,] "instal" "yoshua" "januari" "data" "print"

[8,] "import" "lecun" "softwar" "develop" "function"

[9,] "develop" "schmidhub" "septemb" "posit" "gpu"

[10,] "tutori" "confer" "juli" "vision" "graph"

[11,] "modul" "googl" "april" "engin" "configur"

[12,] "librari" "algorithm" "news" "appli" "program"

[13,] "good" "research" "comment" "scientist" "variabl"

[14,] "question" "hinton" "group" "work" "error"

[15,] "line" "comment" "represent" "scienc" "debug"

[16,] "list" "read" "open" "univers" "time"

[17,] "updat" "intellig" "link" "experi" "modifi"

[18,] "function" "tutori" "novemb" "team" "convolut"

[19,] "learn" "geoffrey" "mila" "imag" "file"

[20,] "featur" "discuss" "list" "applic" "numpi"

Вывод: проанализорованный портал соотвествует своей тематике и может быть рекомендован для дальнейшего изучения специалистами в области образовательных программ для STEM области для углубления своих профессиональных навыков в области глубинного обучения.