**Задание 4. Text mining средствами R.**

Выполнил студент 2 курса

группы 09-715(1)

Санамян Артак.

**Текст задания:**

В этом задании нужно применить методы Text mining к данным, загруженным из Интернет-источников. Для загрузки данных желательно использовать средства R. Можно использовать API социальных сетей, например, Twitter. Желательно загружать тексты на русском языке. Возможные задачи для анализа:

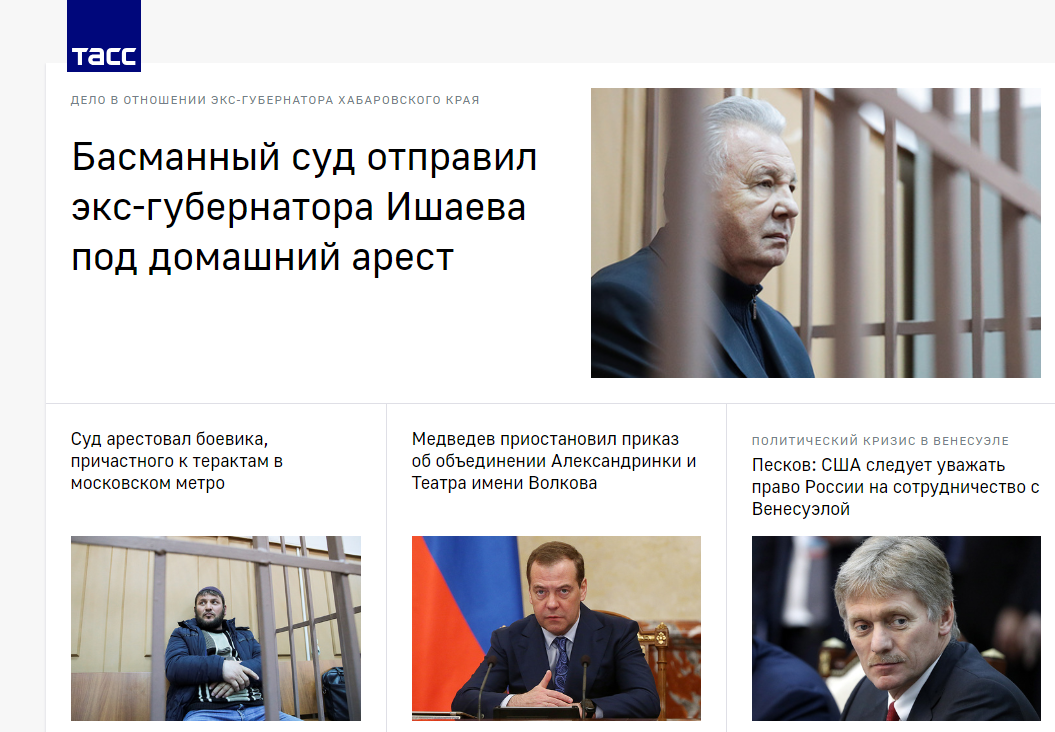
* задача классификации текстов,
* задача кластеризации текстов,
* задача анализа тональности текстов.

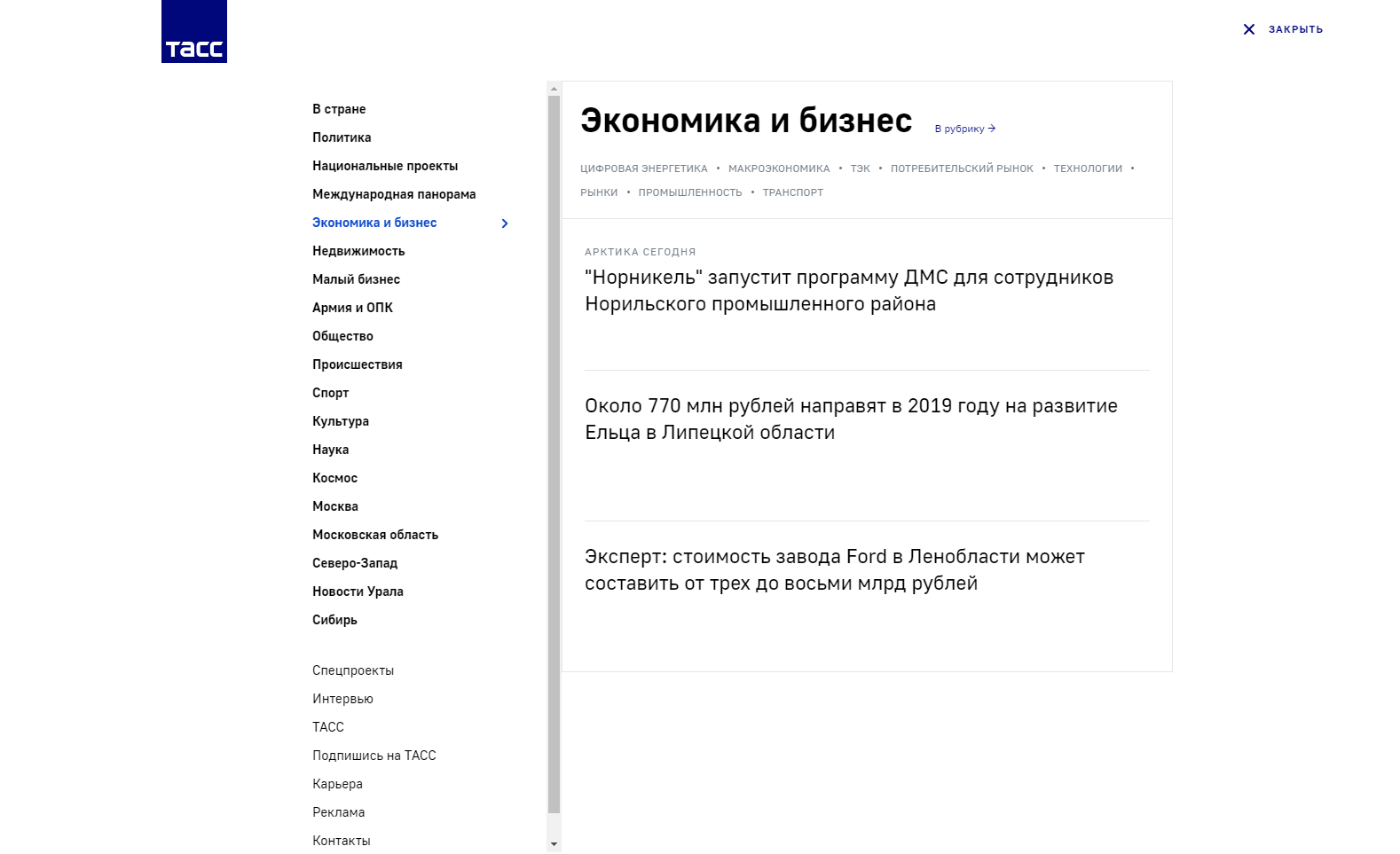
**Описание выполненной работы:**

Рассмотрим задачу классификации новостей по следующим тематикам:

* В стране
* Политика
* Национальные проекты
* Международная панорама
* Экономика и бизнес
* Недвижимость
* Малый бизнес
* Армия и ОПК
* Общество
* Происшествия
* Спорт
* Культура
* Наука
* Космос
* Москва
* Московская область
* Северо-Запад
* Новости Урала
* Сибирь

Новости были взяты с сайта <https://tass.ru>.





Подключим необходимые библиотеки:

> library(RCurl) # основной пакет для веб-скрепинга

> library(XML) # пакет для работы с XML, HTML, Xpath

> library(stringr) # пакет для работы с регулярными выражениями

**# главная страница новостей**

> main = "https://tass.ru/"

**# создаем объект для работы по протоколу SSL**

> signatures = system.file("CurlSSL", cainfo = "cacert.pem", package = "RCurl")

> url = getURL(main, cainfo = signatures, encoding="UTF-8")

**# создаем html-объект**

> webpage <- htmlParse(url, encoding="UTF-8")

Рассмотрим кусок HTML – кода:

|  |
| --- |
| <div class="header-sections\_\_icon"></div>  <div class="header-sections\_\_label">Рубрики</div>  </div>  </div>  </div>  </div>  <button class="header-news-feed-button" data-header\_menu\_feed\_open="data-header\_menu\_feed\_open" ng-hide="ov.variable.newsfeed.isOpened" ng-click="toggleNewsFeed()" ng-class="{'header-news-feed-button\_active': newsFeedLabelActive}"><span ng-cloak="ng-cloak" ng-bind-html="getNewsFeedCounter()"></span></button>  </header>  <div class="search-popup">  <div class="search-popup\_\_container">  <header class="search-popup\_\_header">  <div class="search-popup\_\_logo"></div>  <button class="search-popup\_\_close-button" id="search-btn-close">Закрыть</button>  </header>  <form class="search-popup\_\_search-form">  <div class="search-input">  <div class="search-input\_\_wrapper">  <input class="search-input\_\_input-field" autofocus="autofocus" placeholder="Поиск" data-ng-model="sendParams.searchStr"/><button class="search-input\_\_submit-button" data-ng-click="dispatch()" ng-if="!pending" data-header\_menu\_search\_start="data-header\_menu\_search\_start"></button>  <div class="search-input\_\_spinner" ng-if="pending"></div>  </div>  <ul class="search-input\_\_results-list">  <li class="search-input\_\_results-list-item"><button class="search-input\_\_result-title">Экономика</button></li>  <li class="search-input\_\_results-list-item"><button class="search-input\_\_result-title">Спорт</button></li>  <li class="search-input\_\_results-list-item"><button class="search-input\_\_result-title">Культура</button></li>  </ul>  </div>  </form>  </div>  </div>  <div class="menu" data-ng-controller="menuController" data-ng-class="{'menu\_opened':$root.mainMenuStatus}" menu-directive="menu-directive">  <div class="menu\_\_container">  <header class="menu\_\_header"><a class="menu\_\_logo" href="/" rel="nofollow"></a><button class="menu\_\_close-button" id="close-menu-button" data-ng-click="manageHistoryTrackerForMenu($root.mainMenuStatus=false)" data-all\_category\_close="data-all\_category\_close">Закрыть</button></header>  <div class="menu\_\_content" data-ng-class="{'up':$root.mainMenuStatus}">  <div class="menu\_\_lists-wrapper">  <ul class="menu-sections-list" data-ng-init="init()">  <li class="menu-sections-list-item" data-ng-class="{'menu-sections-list-item\_active': activeSection === 'v-strane'}">  <div class="menu-sections-list\_\_title-wrapper"><a class="menu-sections-list\_\_title" href="/v-strane" data-ng-mouseover="onMouseEnter(6091); activeSection = 'v-strane'" data-all\_category\_choice="data-all\_category\_choice">В стране</a></div>  <div class="section-preview">  <div class="section-preview\_\_content">  <div class="section-preview\_\_header"><span class="section-preview\_\_title">В стране</span><a class="section-preview\_\_link" href="/v-strane" rel="nofollow" data-header\_menu\_categories\_choice="data-header\_menu\_categories\_choice">В рубрику ></a></div>  <ul class="section-preview\_\_subcategories-list"></ul> |

Для получения ссылок на тематики новостей нам нужен тег a с атрибутом class="menu-sections-list\_\_title", перед которым был тег div с атрибутом class='menu-sections-list\_\_title-wrapper':

**# вытащим информацию о тематике новостей**

category <- xpathSApply(webpage, "//div[@class='menu-sections-list\_\_title-wrapper']/a[@class='menu-sections-list\_\_title']", xmlGetAttr, "href");

Уберем из списка категории

* Москва
* Московская область
* Северо-Запад
* Новости Урала
* Сибирь.

Получим 14 тематик:

> category = category[-(15:19)]

> category

[1] "/v-strane" "/politika"

"/nacionalnye-proekty" "/mezhdunarodnaya-panorama"

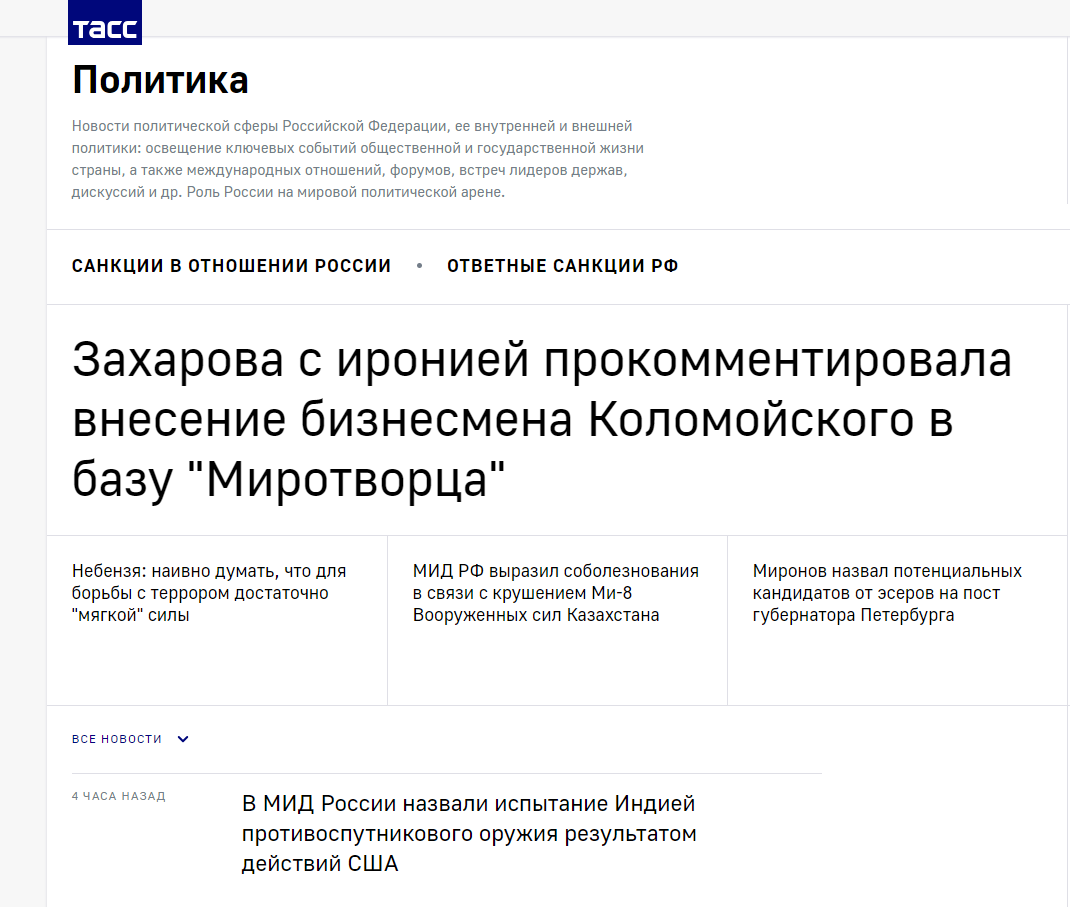
[5] "/ekonomika" "/nedvizhimost"

"/msp" "/armiya-i-opk"

[9] "/obschestvo" "/proisshestviya" "/sport" "/kultura"

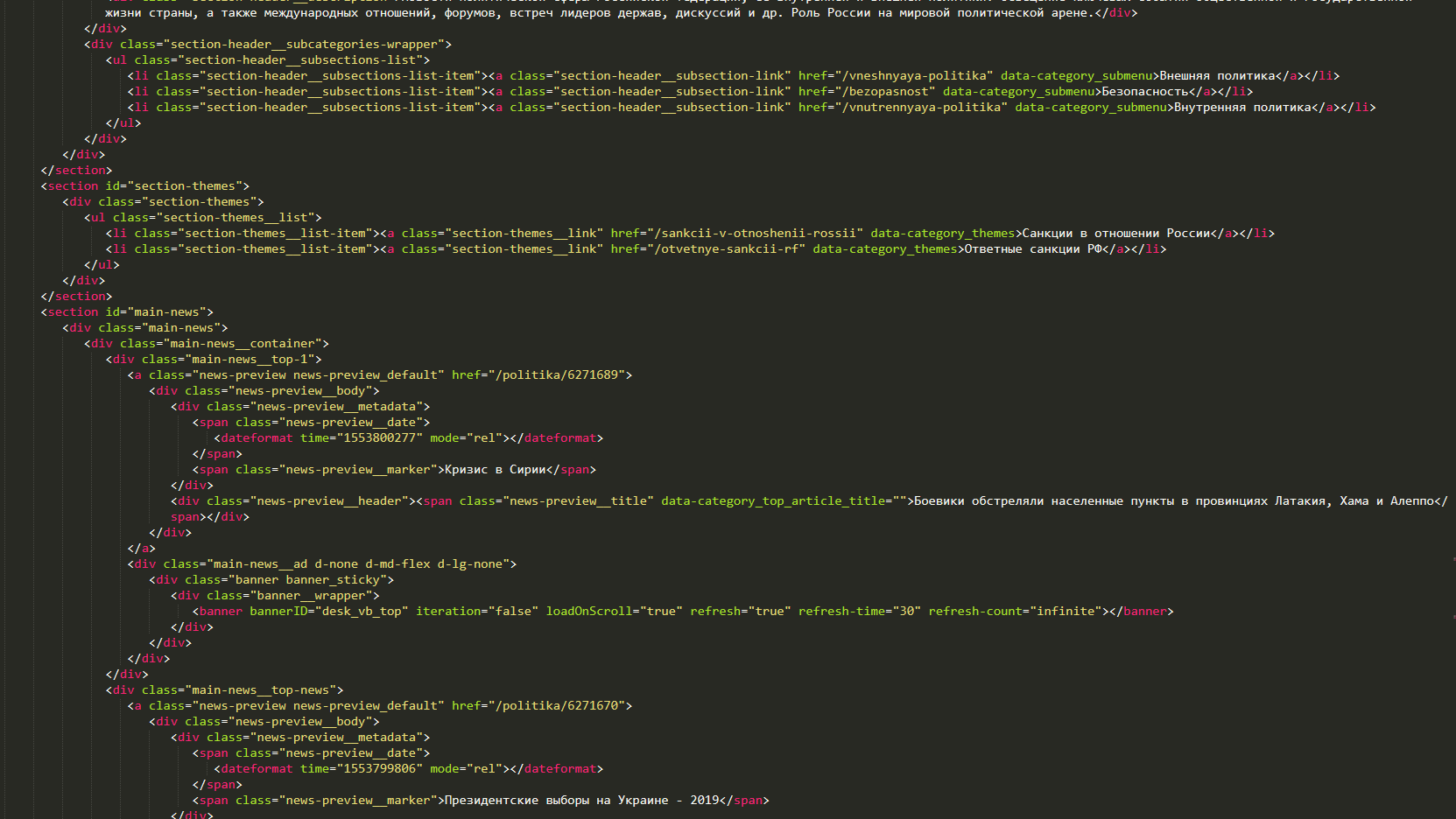
[13] "/nauka" "/kosmos"

В каждой тематике есть своя новостная лента:



Выполним действия для новостей второй тематики (политика).

|  |
| --- |
|  |
|  |

Рассмотрим кусок HTML – кода: 

|  |
| --- |
| <a class="news-preview news-preview\_default" href="/politika/6271689">  <div class="news-preview\_\_body">  <div class="news-preview\_\_metadata">  <span class="news-preview\_\_date">  <dateformat time="1553800277" mode="rel"></dateformat>  </span>  <span class="news-preview\_\_marker">Кризис в Сирии</span>  </div>  <div class="news-preview\_\_header"><span class="news-preview\_\_title" data-category\_top\_article\_title="">Боевики обстреляли населенные пункты в провинциях Латакия, Хама и Алеппо</span></div>  </div>  </a> |

Для получения ссылок на новости текущей тематики нам нужен тег a с атрибутом class='news-preview news-preview\_default'

**#ссылки на новости;**

> ref = character();

**# получаем ссылки на все новости текущей тематики**

**# и сохраняем их в переменную ref**

> # убираем переносы строк:

> tmp <- str\_c(categ\_url, collapse = "")

> # строим HTML-объект:

> tmp <- htmlParse(tmp, encoding="UTF-8")

> # получаем ссылки на все новости текущей тематики

> news\_page <- xpathSApply(tmp, "//a[@class='news-preview news-preview\_default']", xmlGetAttr, "href")

> news\_page

[1] "/kosmos/6271608" "/kosmos/6271160" "/kosmos/6268821" "/kosmos/6268019"

> ref = append(ref, paste0(main1,news\_page))

Рассмотрим кусок HTML – кода:

|  |
| --- |
| <div class="text-content">  <div class="text-block">  <p>МОСКВА, 28 марта. /ТАСС/. Холдинг "Российские космические системы" (РКС, входит в Роскосмос) представил новейшую технологию для термо-видеотелеметрии, которая увеличит надежность ракет-носителей и космических аппаратов, а также поможет решить различные практические задачи на Земле. Об этом в четверг сообщили в пресс-службе РКС.</p>  <p>"Наши специалисты разработали и запатентовали методики применения систем на основе технологий видеотелеметрии. Это очень перспективное направление - предстоит еще много работы, но уже сегодня мы уверены, что можем получить высокоэффективный инструмент контроля ситуации на борту космических аппаратов, в том числе в рамках больших научных проектов по изучению Луны и планет Солнечной системы", - приводят в пресс-службе слова заместителя генерального конструктора РКС Олега Хромова.</p>  <p>Как пояснили в холдинге, новая термо-видеотелеметрическая система позволит определять состояние наблюдаемого объекта по яркости излучения или цветности спектра, которые выделяются из изображения, фиксируемого фоторегистрирующими приборами. Метод обеспечивает контроль температуры крупных узлов и приборов, разогревающихся во время работы.</p>  <p>Установленная на борту космического аппарата система позволяет также фиксировать положение различных объектов и агрегатов, следить за развитием ситуации во время полета. "В дополнение предлагается устанавливать аппаратуру, которая может с высокой точностью фиксировать изменения температуры в разных частях объекта", - отметили в РКС.</p>  <p>Термо-видеотелеметрическая информация, передаваемая системой, будет более полно отражать процессы на борту аппарата, оценивать тепловую нагрузку на узлы и агрегаты. На Земле термо-видеоинформация отображается в цвете на виртуальной модели наблюдаемого объекта.</p>  <p>РКС предлагает использовать эту технологию при создании межорбитальных буксиров, перспективных ракет-носителей и разгонных блоков. Кроме того, на ее базе можно создать линейку систем для наблюдения за сложными и потенциально опасными процессами на Земле - в промышленности, энергетике, авиации и других отраслях. Например, для контроля утечки радиации, мониторинга состояния газопроводов, турбин.</p>  </div>  </div> |

Для получения текста новости нам нужен тег div c атрибутом class="text-block":

> news1\_text <- xpathSApply(news1\_page, "//div[@class='text-block]/p", xmlValue);

**# удаляем из текста знаки переноса строки, табуляции, лишние пробелы:**

> news1\_text = str\_replace\_all(news1\_text, "\n", " ");

> news1\_text = str\_replace\_all(news1\_text, "\t", " ");

> news1\_text = str\_replace\_all(news1\_text, "[ ]{2,}", "");

**# все строки новости записываем в одну переменную full\_text:**

> full\_text = "";

> for(k in 1:length(news1\_text))

+ {

+ full\_text = str\_c(full\_text, paste0(" ", news1\_text[k]));

+ }

**# записываем текст стиха на диск:**

> write(full\_text, paste0("newstext", i, ".txt"), append = TRUE)

Повторим те же действия для загрузки на диск текстов новостей других тематик. Мы имеем на диске 14 файлов вида newstext**i**, где **i** – номер категории:

* Newstext1 – в стране;
* Newstext2 - политика;
* Newstext3 – нац проекты;
* Newstext4 – международная панорама;
* Newstext5 – экономика и бизнес;
* Newstext6 – недвижимость;
* Newstext7 – малый бизнес;
* Newstext8 – армия и опк;
* Newstext9 – общество;
* Newstext10 – происшествия;
* Newstext11 – спорт;
* Newstext12 – культура;
* Newstext13 – наука;
* Newstext14 – космос;

Подготовим тексты для классификации новостей по тематикам.

**# пакет Text mining**

> library(tm)

**# прочитаем новости с диска в переменную news,**

**# а категории, которым принадлежат эти новости, в переменную categ**

> poems = character(0);

> categ = character(0);

**# для классификации выберем новости о политике, о спорте и культуре**

> for(i in c(2, 11, 12))

+ {

+ # читаем новости текущей категории

+ news\_tmp <- readLines(paste0("newstext", i, ".txt"))

+

+ categ\_tmp = rep(name\_category[i], length(news\_tmp))

+

+ news = append(news, news\_tmp)

+ categ = append(categ, categ\_tmp)

+ }

**# сохраним информацию о новостях в data.frame**

> news\_frame = data.frame("news" = news, "category" = categ)

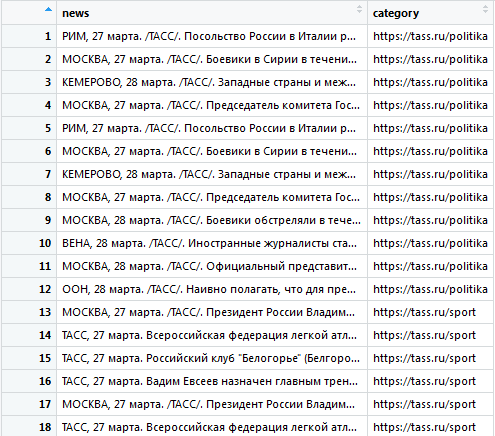
**# общее число новостей**

> nrow(news\_frame)

[1] 36

Посмотрим содержимое news\_frame:

> View(news\_frame)



Заметим, что все новости одной категории расположены подряд. Поэтому перемешаем данные случайный образом:

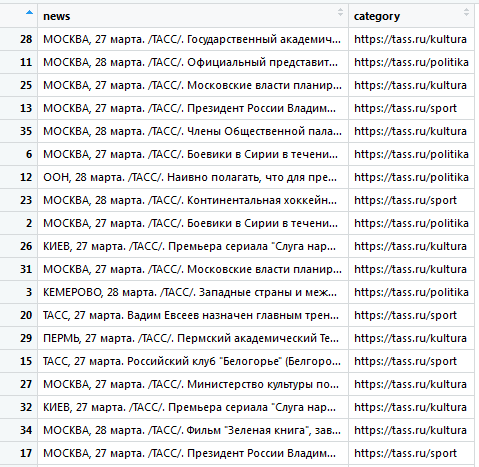
**# перемешаем исходные данные**

> set.seed(0)

> frame = news\_frame[order(runif(nrow(news\_frame))),]

**# просмотрим перемешанные данные**

> View(frame)



**# создаем пустой объект корпус**

> news=list();

> news\_corpus = VCorpus(VectorSource(news))

**# добавляем в корпус информацию о стихах всех тематик:**

> for(i in 1:nrow(frame))

+ {

+ **# добавляем в корпус информацию**

+ tmp\_corpus <- VCorpus(VectorSource(frame[i,]$news))

+ news\_corpus <- c(news\_corpus, tmp\_corpus)

+

+ **# добавляем метаинформацию**

+ meta(news\_corpus[[i]], "category") <- frame[i,]$category

+ meta(news\_corpus[[i]], "language") <- "ru"

+ }

> news\_corpus

<<VCorpus>>

Metadata: corpus specific: 0, document level (indexed): 0

Content: documents: 36

> meta(news\_corpus[[1]])

author : character(0)

datetimestamp: 2019-03-28 20:03:12

description : character(0)

heading : character(0)

id : 1

language : ru

origin : character(0)

category : https://tass.ru/kultura

**# список тематик новостей:**

> meta\_data<- data.frame()

> for (i in 1:NROW(news\_corpus))

+ {

+ meta\_data [i, "category"] <- meta(news\_corpus[[i]], "category")

+ meta\_data [i, "num"] <- i

+ }

> table(as.character(meta\_data[, "category"]))

https://tass.ru/kultura https://tass.ru/politika https://tass.ru/sport

12 12 12

Выполним преобразования данных.

> news\_corpus[[11]]$content

[1] " МОСКВА, 27 марта. /ТАСС/. Московские власти планируют открыть музей в здании Северного речного вокзала после его реставрации. При этом вокзал сохранит свою транспортную функцию, сообщается в среду на информационном портале мэра Москвы. \"Здание Северного речного вокзала после реставрации станет современным транспортным комплексом. Здесь разместится вокзал и музей. В здании оборудуют залы ожидания, билетную кассу, киоск по продаже сувенирной продукции, справочное бюро, буфет, кабинет врача и комнату матери и ребенка. Прилегающую территорию благоустроят\", - говорится в сообщении. Реставрация началась в конце 2018 года. Как сообщал ранее глава Мосгорнаследия Алексей Емельянов, работы будут продолжаться три года. Предполагается, что после реставрации вокзал будет обеспечивать речное сообщение Москвы с Подмосковьем и другими регионами. Планируется воссоздать исторический облик постройки, восстановив по архивным материалам декоративные элементы. В числе прочего специалисты отреставрируют башенные часы XVIII-XIX веков, восстановят настенные панно с изображением шлюзов канала имени Москвы, росписи потолков и наборные мраморные, мозаичные и паркетные полы, отмечают в мэрии. Также реставраторы планируют воссоздать фонари в виде ландышей на крыше. Здание Северного речного вокзала построено в 1930-х гг. по проекту архитекторов Алексея Рухлядева и Владимира Кринского при участии скульптора Ивана Ефимова, художницы Натальи Данько. Оно является объектом культурного наследия регионального значения. В 2010 году комплекс был закрыт из- за аварийного состояния."

**# удалим числа**

> news\_corpus <- tm\_map(news\_corpus, content\_transformer(removeNumbers))

> news\_corpus[[11]]$content

[1] " МОСКВА, марта. /ТАСС/. Московские власти планируют открыть музей в здании Северного речного вокзала после его реставрации. При этом вокзал сохранит свою транспортную функцию, сообщается в среду на информационном портале мэра Москвы. \"Здание Северного речного вокзала после реставрации станет современным транспортным комплексом. Здесь разместится вокзал и музей. В здании оборудуют залы ожидания, билетную кассу, киоск по продаже сувенирной продукции, справочное бюро, буфет, кабинет врача и комнату матери и ребенка. Прилегающую территорию благоустроят\", - говорится в сообщении. Реставрация началась в конце года. Как сообщал ранее глава Мосгорнаследия Алексей Емельянов, работы будут продолжаться три года. Предполагается, что после реставрации вокзал будет обеспечивать речное сообщение Москвы с Подмосковьем и другими регионами. Планируется воссоздать исторический облик постройки, восстановив по архивным материалам декоративные элементы. В числе прочего специалисты отреставрируют башенные часы XVIII-XIX веков, восстановят настенные панно с изображением шлюзов канала имени Москвы, росписи потолков и наборные мраморные, мозаичные и паркетные полы, отмечают в мэрии. Также реставраторы планируют воссоздать фонари в виде ландышей на крыше. Здание Северного речного вокзала построено в -х гг. по проекту архитекторов Алексея Рухлядева и Владимира Кринского при участии скульптора Ивана Ефимова, художницы Натальи Данько. Оно является объектом культурного наследия регионального значения. В году комплекс был закрыт из- за аварийного состояния."

**# заменим на пробелы знаки пунктуации**

> news\_corpus <- tm\_map(news\_corpus,

+ content\_transformer(str\_replace\_all),

+ pattern = "[[:punct:]]", replacement = " ")

> news\_corpus[[11]]$content

[1] " МОСКВА марта ТАСС Московские власти планируют открыть музей в здании Северного речного вокзала после его реставрации При этом вокзал сохранит свою транспортную функцию сообщается в среду на информационном портале мэра Москвы Здание Северного речного вокзала после реставрации станет современным транспортным комплексом Здесь разместится вокзал и музей В здании оборудуют залы ожидания билетную кассу киоск по продаже сувенирной продукции справочное бюро буфет кабинет врача и комнату матери и ребенка Прилегающую территорию благоустроят говорится в сообщении Реставрация началась в конце года Как сообщал ранее глава Мосгорнаследия Алексей Емельянов работы будут продолжаться три года Предполагается что после реставрации вокзал будет обеспечивать речное сообщение Москвы с Подмосковьем и другими регионами Планируется воссоздать исторический облик постройки восстановив по архивным материалам декоративные элементы В числе прочего специалисты отреставрируют башенные часы XVIII XIX веков восстановят настенные панно с изображением шлюзов канала имени Москвы росписи потолков и наборные мраморные мозаичные и паркетные полы отмечают в мэрии Также реставраторы планируют воссоздать фонари в виде ландышей на крыше Здание Северного речного вокзала построено в х гг по проекту архитекторов Алексея Рухлядева и Владимира Кринского при участии скульптора Ивана Ефимова художницы Натальи Данько Оно является объектом культурного наследия регионального значения В году комплекс был закрыт из за аварийного состояния "

**# преобразуем к нижнему регистру**

> news\_corpus <- tm\_map(news\_corpus, content\_transformer(tolower))

> news\_corpus[[11]]$content

[1] " москва марта тасс московские власти планируют открыть музей в здании северного речного вокзала после его реставрации при этом вокзал сохранит свою транспортную функцию сообщается в среду на информационном портале мэра москвы здание северного речного вокзала после реставрации станет современным транспортным комплексом здесь разместится вокзал и музей в здании оборудуют залы ожидания билетную кассу киоск по продаже сувенирной продукции справочное бюро буфет кабинет врача и комнату матери и ребенка прилегающую территорию благоустроят говорится в сообщении реставрация началась в конце года как сообщал ранее глава мосгорнаследия алексей емельянов работы будут продолжаться три года предполагается что после реставрации вокзал будет обеспечивать речное сообщение москвы с подмосковьем и другими регионами планируется воссоздать исторический облик постройки восстановив по архивным материалам декоративные элементы в числе прочего специалисты отреставрируют башенные часы xviii xix веков восстановят настенные панно с изображением шлюзов канала имени москвы росписи потолков и наборные мраморные мозаичные и паркетные полы отмечают в мэрии также реставраторы планируют воссоздать фонари в виде ландышей на крыше здание северного речного вокзала построено в х гг по проекту архитекторов алексея рухлядева и владимира кринского при участии скульптора ивана ефимова художницы натальи данько оно является объектом культурного наследия регионального значения в году комплекс был закрыт из за аварийного состояния "

**# удаляем стоп-слова**

> news\_corpus <- tm\_map(news\_corpus, content\_transformer(removeWords), words = stopwords("ru"))

> news\_corpus[[11]]$content

[1] " москва марта тасс московские власти планируют открыть музей здании северного речного вокзала реставрации вокзал сохранит транспортную функцию сообщается среду информационном портале мэра москвы здание северного речного вокзала реставрации станет современным транспортным комплексом разместится вокзал музей здании оборудуют залы ожидания билетную кассу киоск продаже сувенирной продукции справочное бюро буфет кабинет врача комнату матери ребенка прилегающую территорию благоустроят говорится сообщении реставрация началась конце года сообщал ранее глава мосгорнаследия алексей емельянов работы будут продолжаться года предполагается реставрации вокзал обеспечивать речное сообщение москвы подмосковьем другими регионами планируется воссоздать исторический облик постройки восстановив архивным материалам декоративные элементы числе прочего специалисты отреставрируют башенные часы xviii xix веков восстановят настенные панно изображением шлюзов канала имени москвы росписи потолков наборные мраморные мозаичные паркетные полы отмечают мэрии также реставраторы планируют воссоздать фонари виде ландышей крыше здание северного речного вокзала построено х гг проекту архитекторов алексея рухлядева владимира кринского участии скульптора ивана ефимова художницы натальи данько оно является объектом культурного наследия регионального значения году комплекс закрыт аварийного состояния "

**# проводим стемминг**

> for(i in 1:NROW(news\_corpus))

+ news\_corpus[[i]]$content <- stemDocument(news\_corpus[[i]]$content, language = "russian")

> news\_corpus <- tm\_map(news\_corpus, content\_transformer(removeWords), words = 'котор')

> news\_corpus[[11]]$content

[1] "москв март тасс московск власт планир откр муз здан северн речн вокза реставрац вокза сохран транспортн функц сообща сред информацион портал мэр москв здан северн речн вокза реставрац станет современ транспортн комплекс размест вокза муз здан оборуд зал ожидан билетн касс киоск продаж сувенирн продукц справочн бюр буфет кабинет врач комнат матер ребенк прилега территор благоустро говор сообщен реставрац нача конц год сообща ран глав мосгорнаслед алекс емельян работ будут продолжа год предполага реставрац вокза обеспечива речн сообщен москв подмосков друг регион планир воссозда историческ облик постройк восстанов архивн материал декоративн элемент числ проч специалист отреставрир башен час xviii xix век восстанов настен пан изображен шлюз кана имен москв роспис потолк наборн мраморн мозаичн паркетн пол отмеча мэр такж реставратор планир воссозда фонар вид ландыш крыш здан северн речн вокза постро х гг проект архитектор алексе рухлядев владимир кринск участ скульптор ива ефимов художниц натал даньк он явля объект культурн наслед региональн значен год комплекс закр аварийн состоян"

**# создадим облако слов**

> library(wordcloud)

> wordcloud(news\_corpus, random.order=F, max.words=40)



**# создаем матрицу терминов-документов**

> tdm <- TermDocumentMatrix(news\_corpus)

> tdm

<<TermDocumentMatrix (terms: 2053, documents: 36)>>

Non-/sparse entries: 4768/69140

Sparsity : 94%

Maximal term length: 19

Weighting : term frequency (tf)

**# создаем матрицу документов-терминов**

> dtm <- DocumentTermMatrix(news\_corpus)

> dtm

<<DocumentTermMatrix (documents: 36, terms: 2053)>>

Non-/sparse entries: 4768/69140

Sparsity : 94%

Maximal term length: 19

Weighting : term frequency (tf)

> dtm <- removeSparseTerms(dtm, 0.97)

> dtm

<<DocumentTermMatrix (documents: 36, terms: 1187)>>

Non-/sparse entries: 3902/38830

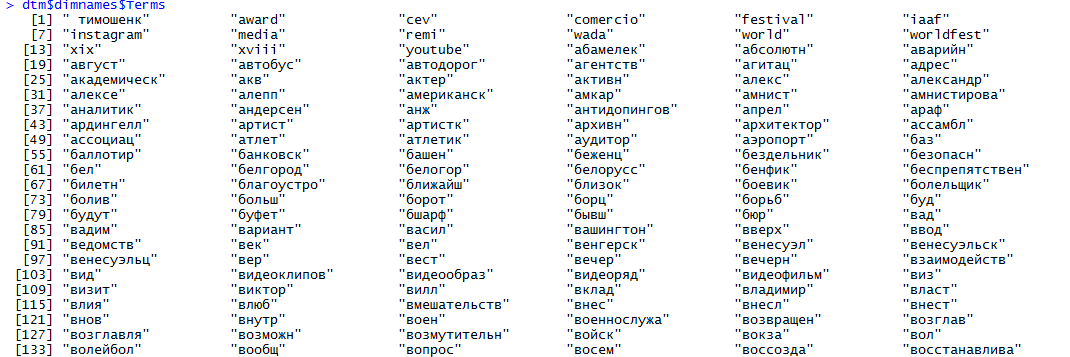
Sparsity : 91%

Maximal term length: 18

Weighting : term frequency (tf)

**# список терминов**

> dtm$dimnames$Terms



**# дополнительный пакет для работы с текстами**

> library(RTextTools)

**# массив меток – тематики новостей**

> org\_labels<-meta\_data[, "category"]

> count\_data <- length(org\_labels)

> count\_data

[1] 36

**# разделим исходные данные на тестовую и обучающую выборки**

**# число новостей обучающей выборки (70% от исходных данных)**

> count\_train = round(count\_data\*0.7)

> count\_train

[1] 25

**# для применения методов классификации создаем контейнер**

> container <- create\_container(

+ dtm,

+ labels = org\_labels,

+ trainSize = 1:count\_train,

+ testSize = (count\_train+1):count\_data,

+ virgin = FALSE

+ )

**# обучение моделей**

**# Support vector machine**

> svm\_model <- train\_model(container, "SVM")

**# Random forest**

> tree\_model <- train\_model(container, "TREE")

**# Maximal entropy**

> maxent\_model <- train\_model(container, "MAXENT")

**# запускаем классификацию**

> svm\_out <- classify\_model(container, svm\_model)

> tree\_out <- classify\_model(container, tree\_model)

> maxent\_out <- classify\_model(container, maxent\_model)

**# создаем общую матрицу меток**

> labels\_out <- data.frame(

+ correct\_label = org\_labels[(count\_train+1):count\_data],

+ svm = as.character(svm\_out[,1]),

+ tree = as.character(tree\_out[,1]),

+ maxent = as.character(maxent\_out[,1]),

+ stringsAsFactors = F)

**# SVM performance**

> table(labels\_out[,1] == labels\_out[,2])

FALSE TRUE

2 9

**# Random forest performance**

> table(labels\_out[,1] == labels\_out[,3])

FALSE TRUE

6 5

**# Maximum entropy performance**

> table(labels\_out[,1] == labels\_out[,4])

FALSE TRUE

2 9

Таким образом, метод SVM допустил 2 ошибки из 11 (точность метода составляет 81,8%); метод Random forest допустил 6 ошибок из 11 (точность метода составляет 55,5%); метод Maximum entropy допустил 2 ошибки из 11 (точность метода составляет 81,8%).

Попробуем уменьшить разреженность матрицы документов-терминов:

> dtm <- removeSparseTerms(dtm, 0.9)

> dtm

<<DocumentTermMatrix (documents: 36, terms: 289)>>

Non-/sparse entries: 1887/8517

Sparsity : 82%

Maximal term length: 13

Weighting : term frequency (tf)

Было 1187 терминов, стало 289.

**# список терминов**

> dtm$dimnames$Terms

[105] "лин" "лиц" "лондон" "лучш" "люб" "люд" "максимальн" "мар"

[113] "март" "матч" "международн" "мест" "мид" "министр" "мир" "млн"

[121] "мнен" "мобильн" "могут" "москв" "московск" "музыкальн" "назва" "налаж"

[129] "нам" "напомн" "народ" "населен" "наход" "национальн" "нача" "начина"

[137] "наш" "необходим" "нов" "ноябр" "нужн" "обещан" "обладател" "област"

[145] "образ" "обществен" "объект" "одн" "октябр" "олимпийск" "оппозиц" "организац"

[153] "оста" "откр" "отмет" "отмеча" "отношен" "официальн" "очен" "палат"

[161] "перв" "планир" "площадк" "поддержк" "подобн" "подписа" "подход" "подчеркнул"

В этом случае точность методов классификации уменьшится:

**# SVM performance**

> table(labels\_out[,1] == labels\_out[,2])

FALSE TRUE

3 8

**# Random forest performance**

> table(labels\_out[,1] == labels\_out[,3])

FALSE TRUE

6 5

**# Maximum entropy performance**

> table(labels\_out[,1] == labels\_out[,4])

FALSE TRUE

2 9

Как видим, точность метода SVM понизилась до 72%, точность метода Random forest не изменилась. Дальнейшее изменение разреженности матрицы не привело к увеличению точности классификации.

Попробуем запустить классификацию новостей для 5 тематик:

* политика;
* спорт;
* культура;
* общество;
* наука.

**# общее число новостей**

> nrow(news\_frame)

[1] 60

> table(as.character(meta\_data[, "category"]))

https://tass.ru/kultura 12

<https://tass.ru/nauka> 12

https://tass.ru/obschestvo 12

https://tass.ru/politika 12

https://tass.ru/sport 12

> dtm <- DocumentTermMatrix(news\_corpus)

> dtm

<<DocumentTermMatrix (documents: 60, terms: 3147)>>

Non-/sparse entries: 8415/180405

Sparsity : 96%

Maximal term length: 19

Weighting : term frequency (tf)

> dtm <- removeSparseTerms(dtm, 0.97)

> dtm

<<DocumentTermMatrix (documents: 60, terms: 1898)>>

Non-/sparse entries: 7166/106714

Sparsity : 94%

Maximal term length: 19

Weighting : term frequency (tf)

**# SVM performance**

> table(labels\_out[,1] == labels\_out[,2])

FALSE TRUE

3 15

**# Random forest performance**

> table(labels\_out[,1] == labels\_out[,3])

FALSE TRUE

8 10

**# Maximum entropy performance**

> table(labels\_out[,1] == labels\_out[,4])

FALSE TRUE

3 15

Метод SVM допустил 3 ошибок из 18 (точность метода составляет 83,3%); метод Random forest допустил 8 ошибок из 18 (точность метода составляет 55,5%); метод Maximum entropy допустил 3 ошибок из 18 (точность метода составляет 83,3%).

**Код программы:**

# назначим рабочий директорий

setwd("C:/Users/admin.WIN-OSD1NEIENB0/Documents/Rprojects/task4/tass")

library(RCurl) # основной пакет для веб-скрепинга

library(XML) # пакет для работы с XML, HTML, Xpath

library(stringr) # пакет для работы с регулярными выражениями

# новости, сми

main = "https://tass.ru/"

# создаем объект для работы по протоколу SSL

signatures = system.file("CurlSSL", cainfo = "cacert.pem", package = "RCurl")

url = getURL(main, cainfo = signatures, encoding="UTF-8")

write(url, 'tassurl.txt')

# создаем html-объект

webpage <- htmlParse(url, encoding="UTF-8")

webpage

# вытащим информацию о тематике новостей

category <- xpathSApply(webpage, "//div[@class='menu-sections-list\_\_title-wrapper']/a[@class='menu-sections-list\_\_title']", xmlGetAttr, "href");

category = category[-(15:19)]

category

main1 = substr(main, 1, nchar(main)-1)

for(i in 1:length(category))

{

category[i] = paste0(main1, category[i])

categ\_url=getURL(category[i], cainfo = signatures, encoding="UTF-8")

# загружаем все страницы на диск:

write(categ\_url, paste0('Categ/categ', i, '.txt'))

catpage <- htmlParse(categ\_url, encoding="UTF-8")

#catpage

#ссылки на новости;

ref = character();

# убираем переносы строк:

tmp <- str\_c(categ\_url, collapse = "")

# строим HTML-объект:

tmp <- htmlParse(tmp, encoding="UTF-8")

# получаем ссылки на все новости текущей тематики

news\_page <- xpathSApply(tmp, "//a[@class='news-preview news-preview\_default']", xmlGetAttr, "href")

news\_page

ref = append(ref, paste0(main1,news\_page))

# записываем на диск тексты всех новостей текущей тематики

for(j in 1:length(ref))

{

print(j)

news1\_url= getURL(ref[j], cainfo = signatures, encoding="UTF-8");

# write(news1\_url, paste0('News/news', i, '\_', j, '.txt'))

news1\_page <- htmlParse(news1\_url, encoding="UTF-8")

news1\_text <- xpathSApply(news1\_page, "//div[@class='text-block']/p", xmlValue)

# удаляем из текста знаки переноса строки, табуляции, лишние пробелы:

news1\_text = str\_replace\_all(news1\_text, "\n", " ");

news1\_text = str\_replace\_all(news1\_text, "\t", " ");

news1\_text = str\_replace\_all(news1\_text, "[ ]{2,}", "");

# все строки новости записываем в одну переменную full\_text:

full\_text = "";

for(k in 1:length(news1\_text))

{

full\_text = str\_c(full\_text, paste0(" ", news1\_text[k]));

}

# записываем текст новости на диск:

write(full\_text, paste0("Newstext/newstext", i, ".txt"), append = TRUE)

}

}

# пакет Text mining

library(tm)

# получим названия тематик новостей

#name\_category <- xpathSApply(webpage, "//div[@class='menu-menyu-1-container']/ul[@class='menu']/li[@class='menu-item menu-item-has-children']/ul[@class='sub-menu']/li/a", xmlValue);

name\_category = category

# прочитаем новости с диска в переменную news,

# а категории, которым принадлежат эти новости, в переменную categ

news = character(0);

categ = character(0);

# для классификации выберем новости о политике, спорте и культуре (обществе 9, науке 13)

for(i in c(2, 11, 12)) #9, 13

{

# читаем новости текущей категории

news\_tmp <- readLines(paste0("Newstext/newstext", i, ".txt"))

categ\_tmp = rep(name\_category[i], length(news\_tmp))

news = append(news, news\_tmp)

categ = append(categ, categ\_tmp)

}

# сохраним информацию о новостях в data.frame

news\_frame = data.frame("news" = news, "category" = categ)

# общее число новостей

nrow(news\_frame)

# перемешаем исходные данные

set.seed(0)

frame = news\_frame[order(runif(nrow(news\_frame))),]

# создаем пустой объект корпус

news=list();

news\_corpus = VCorpus(VectorSource(news))

# добавляем в корпус информацию о новостях всех тематик:

for(i in 1:nrow(frame))

{

# добавляем в корпус информацию

tmp\_corpus <- VCorpus(VectorSource(frame[i,]$news))

news\_corpus <- c(news\_corpus, tmp\_corpus)

# добавляем метаинформацию

meta(news\_corpus[[i]], "category") <- frame[i,]$category

meta(news\_corpus[[i]], "language") <- "ru"

}

news\_corpus

meta(news\_corpus[[1]])

# список тематик новостей:

meta\_data<- data.frame()

for (i in 1:NROW(news\_corpus))

{

meta\_data [i, "category"] <- meta(news\_corpus[[i]], "category")

meta\_data [i, "num"] <- i

}

table(as.character(meta\_data[, "category"]))

news\_corpus[[11]]$content

# удалим числа

news\_corpus <- tm\_map(news\_corpus, content\_transformer(removeNumbers))

news\_corpus[[11]]$content

# заменим на пробелы знаки пунктуации

news\_corpus <- tm\_map(news\_corpus,

content\_transformer(str\_replace\_all),

pattern = "[[:punct:]]", replacement = " ")

news\_corpus[[11]]$content

# преобразуем к нижнему регистру

news\_corpus <- tm\_map(news\_corpus, content\_transformer(tolower))

news\_corpus[[11]]$content

# удаляем стоп-слова

news\_corpus <- tm\_map(news\_corpus, content\_transformer(removeWords), words = stopwords("ru"))

news\_corpus[[11]]$content

# проводим стемминг

for(i in 1:NROW(news\_corpus))

news\_corpus[[i]]$content <- stemDocument(news\_corpus[[i]]$content, language = "russian")

news\_corpus <- tm\_map(news\_corpus, content\_transformer(removeWords), words = 'котор')

news\_corpus[[11]]$content

# создадим облако слов

library(wordcloud)

wordcloud(news\_corpus, random.order=F, max.words=40)

# создаем матрицу терминов-документов

tdm <- TermDocumentMatrix(news\_corpus)

tdm

# создаем матрицу документов-терминов

dtm <- DocumentTermMatrix(news\_corpus)

dtm

dtm <- removeSparseTerms(dtm, 0.97)

dtm

# список терминов

dtm$dimnames$Terms

# дополнительный пакет для работы с текстами

library(RTextTools)

# массив меток – тематики новостей

org\_labels<-meta\_data[, "category"]

count\_data <- length(org\_labels)

count\_data

# разделим исходные данные на тестовую и обучающую выборки

# число новостей обучающей выборки (70% от исходных данных)

count\_train = round(count\_data\*0.7)

count\_train

# для применения методов классификации создаем контейнер

container <- create\_container(

dtm,

labels = org\_labels,

trainSize = 1:count\_train,

testSize = (count\_train+1):count\_data,

virgin = FALSE

)

container

# проверим репрезентативность выборки

round(prop.table(table(frame[2]))\*100, digits = 1)

train\_data = frame[1:count\_train,]

round(prop.table(table(train\_data[2]))\*100, digits = 1)

# обучение моделей

# Support vector machine

svm\_model <- train\_model(container, "SVM")

# Random forest

tree\_model <- train\_model(container, "TREE")

# Maximal entropy

maxent\_model <- train\_model(container, "MAXENT")

# запускаем классификацию

svm\_out <- classify\_model(container, svm\_model)

tree\_out <- classify\_model(container, tree\_model)

maxent\_out <- classify\_model(container, maxent\_model)

# создаем общую матрицу меток

labels\_out <- data.frame(

correct\_label = org\_labels[(count\_train+1):count\_data],

svm = as.character(svm\_out[,1]),

tree = as.character(tree\_out[,1]),

maxent = as.character(maxent\_out[,1]),

stringsAsFactors = F)

# SVM performance

table(labels\_out[,1] == labels\_out[,2])

# Random forest performance

table(labels\_out[,1] == labels\_out[,3])

# Maximum entropy performance

table(labels\_out[,1] == labels\_out[,4])