**ПРЕДОБРАБОТКА И РАЗМЕТКА**

**Шаг 1.** Разбить текст на предложения, пометить начало и конец каждого предложения.

**Шаг 2.** Разбить каждое предложение на клауземы, заменить знаки препинания внутри предложения, кавычки и союзы специальными символами.

**Шаг 3.** Разбить текст на слова, используя пробелы.

**Шаг 4.** Подсчитать количество слов в тексте.

**Шаг 5.** Пометить слова, набранные прописными буквами.

**Шаг 6.** Заменить все прописные буквы строчными.

**Шаг 7.** Подвергнуть все слова лемматизации.

Изначально данный алгоритм использовался для системы определения тональностей текстов для русского языка, однако, немного модифицировав его, мы смогли применить его в нашем исследовании.

Условные обозначения и специальные обозначения для работы с текстом приведены в следующей таблице.

Табл. 1 Служебные символы для корректирующего лексикона

|  |  |  |
| --- | --- | --- |
| **Корректирующий лексикон** | **Примеры** | **Служебные символы** |
| Модификаторы полярности | *No, not, neither, never, none, nobody, nothing, nowhere* | ALT |
|  | *without* | WT |
|  | ? ?! (вопросительный знак, вопросительный знак с восклицательным знаком) | QM |
|  | “ “ (кавычки) | Q |
| Усилители полярности | *Very, absolutely, completely, ever* | INC |
|  | Слово, набранное прописными буквами | CAP |
|  | ! (восклицательный знак) | EM |
| Антимодификаторы полярности | *So, such* | TH |
| Прочие символы | Предложение | S |
|  | Фрагмент | F |
|  | союз *and* / *or* | CJ |
|  | , : ; – (запятая, двоеточие, точка с запятой, тире) | Z |
|  | Слово, не входящее в лексикон | *w* |

**ПРИМЕР**

Для большей наглядности рассмотрим работу данного алгоритма на примере комментария с сайта reddit.com/r/politics*:*

*“The CIA has been incompetent from its inception. The roster of incompetence includes subversion operations that cost the lives of hundreds of agents and accomplished nothing; CIA-managed coups that backfired, the Bay of Pigs and many others. Even operations that "succeeded" were pyrrhic”.*

**Шаг 1.** Разбить текст на предложения, пометить начало и конец каждого предложения:

* The CIA has been incompetent from its inception.
* The roster of incompetence includes subversion operations that cost the lives of hundreds of agents and accomplished nothing; CIA-managed coups that backfired, the Bay of Pigs and many others.
* Even operations that "succeeded" were pyrrhic.

**Шаг 2.** Разбить каждое предложение на клауземы, заменить знаки препинания внутри предложения, кавычки и союзы специальными символами:

* The CIA
* has been incompetent
* from its inception<P>
* The roster of incompetence
* includes subversion operations
* that cost the lives of hundreds of agents
* CJ accomplished nothing<Z>
* CIA-managed coups
* that backfired<Z>
* the Bay of Pigs
* CJ many others<P>
* Even operations that <Q>succeeded<Q>
* were pyrrhic<P**>**

**Шаг 3.** Разбить текст на слова, используя пробелы:

* The
* CIA
* has
* been
* incompetent
* from
* its
* inception<P>
* The
* roster
* of
* incompetence
* includes
* subversion
* operations
* that
* cost
* the
* lives
* of
* hundreds
* of
* agents
* CJ
* accomplished
* nothing<Z>
* CIA-managed
* coups
* that
* backfired<Z>
* the
* Bay
* of
* Pigs
* CJ
* many
* others<P>
* Even
* operations
* that
* <Q>succeeded<Q>
* were pyrrhic
* <P**>**

**Шаг 4.** Подсчитать количество слов в тексте: 43 слова.

**Шаг 5.** Пометить слова, набранные прописными буквами: CIA > CIA<CAP>

**Шаг 6.** Заменить все прописные буквы строчными:

* the
* CIA<CAP>
* has
* been
* incompetent
* from
* its
* inception<P>
* the
* roster
* of
* incompetence
* includes
* subversion
* operations
* that
* cost
* the
* lives
* of
* hundreds
* of
* agents
* CJ
* accomplished
* nothing<Z>
* CIA-managed
* coups
* that
* backfired<Z>
* the
* bay
* of
* pigs
* CJ
* many
* others<P>
* even
* operations
* that
* <Q>succeeded<Q>
* were pyrrhic
* <P**>**

**Шаг 7.** Подвергнуть все слова лемматизации:

* the
* CIA<CAP>
* have
* been
* incompetent
* from
* its
* inception<P>
* the
* roster
* of
* incompetence
* include
* subversion
* operation
* that
* cost
* the
* live
* of
* hundred
* of
* agent
* CJ
* accomplish
* nothing<Z>
* CIA-managed
* coup
* that
* backfire<Z>
* the
* bay
* of
* pig
* CJ
* many
* other<P>
* even
* operation
* that
* <Q>succeed<Q>
* pyrrhic
* <P>

Далее рассмотрим правила для корректировки черновой полярности, представленные ниже.

**ПРАВИЛА**

**Правило 1.1.** <ALT>(<POS>){n}→<NEG>{n}

Если в промежутке от начала предложения или знака препинания или союза and*, or* до следующего знака препинания или союза *and, or* имеется модификатор полярности, то положительная полярность всех слов, входящих в оценочный лексикон, в данном промежутке изменяется на отрицательную. Порядок указанных элементов (модификатор, слово с полярностью, любое другое слово) значения не имеет.

*‘It’s not good for our democracy’: Calls grow for federal officers to shed military-style uniforms* Count= -1.

**Правило 1.2.** <ALT>(<NEG>){n}→<POS>{n}

Если в промежутке от начала предложения или знака препинания или союза and*, or* до следующего знака препинания или союза and*, or* имеется модификатор полярности, то отрицательная полярность всех слов, входящих в оценочный лексикон, в данном промежутке изменяется на положительную. Порядок указанных элементов (модификатор, слово с полярностью, любое другое слово) значения не имеет.

*Facebook, finally, does something not terrible.* Count=+1*.*

**Правило 2.1.** (<INC><NEG>) →<NEG><NEG>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеются слова, входящие в отрицательный лексикон, а также имеется усилитель, то каждый усилитель засчитывается за одно слово отрицательного лексикона. Порядок указанных элементов (усилитель, слово с полярностью, любое другое слово) значения не имеет.

*It was a very unprofessional memorandum.* Count= -2.

**Правило 2.2.** (<INC><POS>) →<POS><POS>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеются слова, входящие в положительный лексикон, а также имеется усилитель, то каждый усилитель засчитывается за одно слово положительного лексикона. Порядок указанных элементов (усилитель, слово с полярностью, любое другое слово) значения не имеет.

*Having CU overturned would be absolutely GREAT, but there are so many things that prevent Congress from accurately representing us that I wouldn't even know where to go after that....* Count=+2 (+1 за использование заглавных букв в слове положительного лексикона).

**Правило 3.1.** (<TH><ALT><NEG>)→<NEG>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеется модификатор полярности (ALT) и антимодификатор полярности (TH), то полярность слов в данном промежутке не изменяется, т.е. Правило 1 не применяется. Порядок указанных элементов (модификатор, антимодификатор, слово с полярностью, любое другое слово) значения не имеет.

*Never have we seen such chaos and corruption.* Count= -1.

**Правило 3.2.** (<TH><ALT><POS>)→ <POS>

Если в промежутке от начала предложения или знака препинания до следующего знака препинания имеется модификатор полярности (ALT) и антимодификатор полярности (TH), то полярность слов в данном промежутке не изменяется, т.е. Правило 1 не применяется. Порядок указанных элементов (модификатор, антимодификатор, слово с полярностью, любое другое слово) значения не имеет.

*Lee also noted he has not seen such genuine support for real change from such an enormous, diverse group since the 1960s.* Count= +1.

**Правило 4.** <POS>{1}<QM>→<NEG>

Если предложение состоит из одного слова с положительной полярностью и заканчивается ? или ?!, то полярность данного слова изменяется на отрицательную.

*Trustworthy?* Count= -1.

**Правило 5.** <QM>→<NEG>

Если предложение заканчивается вопросительным знаком или вопросительным знаком с восклицательным знаком, при этом в предложении нет слов, входящих в оценочный лексикон, то полярность данного предложения равна -1.

*Is this a joke?* Count= -1.

**Правило 6.** <Q><POS><Q>→<NEG>

Если слово с положительной полярностью заключено в кавычки, то полярность данного слова изменяется на отрицательную. Кавычки являются модификатором полярности.

*Even operations that "succeeded" were pyrrhic*. Count= -1.

**Правило 7.1.** <WT>(*w|*INC){0,}<POS>→<NEG>

Если перед словом с положительной полярностью имеется слово *without*, то полярность данного слова изменяется на отрицательную. Слово *without* является модификатором полярности. Вместе со словом *without* и со словом с полярностью может быть усилитель (*Very, absolutely, completely, ever*).

*...powers have been without justification, and after two-hundred plus years in practice, without prudence.* Count= -1.

**Правило 7.2.** <WT>(*w|*INC){0,}<NEG>→<POS>

Если перед словом с отрицательной полярностью имеется слово *without*, то полярность данного слова изменяется на положительную. Слово *without* является модификатором полярности. Вместе со словом *without* и со словом с полярностью может быть усилитель (*Very, absolutely, completely, ever*).

*They are being fed Democrat talking points, and they play them without hesitation or research.* Count=+1.

*...and the maximum number of people infected before states can open up completely without endangering the public.* Count=+2.

**Правило 8.1.** <POS><EM>→<POS><POS>

Если предложение заканчивается восклицательным знаком, а счет предложения положительный, то восклицательный знак приравнивается к одному слову с положительной полярностью.

*Great work again, Terrence has been on a roll this week!* Count=2.

**Правило 8.2.** <NEG><EM>→<NEG><NEG>

Если предложение заканчивается восклицательным знаком, или вопросительным знаком, или вопросительным знаком с восклицательным знаком, а счет предложения отрицательный, то восклицательный знак, или вопросительный знак, или вопросительный знак с восклицательным знаком приравнивается к одному слову с отрицательной полярностью.

*"They did a terrible job!"* Count= -2.

*FEDS, Y U no listen?!* Count= -2.

*Awful job draining the swamp, huh?* Count=-2.

**Правило 9.1.** <POS><CAP>→<POS><POS>

Если в предложении имеются слова из прописных букв, а счет предложения положительный, то каждое слово из прописных букв приравнивается к одному слову с положительной полярностью.

*Is it just me or is it AWESOME to go the whole week without hearing a WORD about Hillary Clinton?!* Count=+2 (+1 за знаки ?!).

**Правило 9.2.** <NEG><CAP>→<NEG><NEG>

Если в предложении имеются слова из прописных букв, а счет предложения отрицательный, то каждое слово из прописных букв приравнивается к одному слову с отрицательной полярностью.

*I have NO idea who to vote for.* Count=-2.

Таким образом, вновь рассмотрим пример *“The CIA has been incompetent from its inception. The roster of incompetence includes subversion operations that cost the lives of hundreds of agents and accomplished nothing; CIA-managed coups that backfired, the Bay of Pigs and many others. Even operations that "succeeded" were pyrrhic”.*  Слова *incompetent* и *incompetence* относятся к оценочному лексикону, с отрицательной ценностью: count= -2. Далее, слово *accomplish* с положительной ценностью, однако при нём стоит модификатор полярности, значит применяется правило 1.1, следовательно, count=-1. Наконец, слово *succeeded* так же имеет положительную ценность, но взято в кавычки. Применяем правило 6, count=-1. В итоге, оценочная ценность данного комментария равна -4 => комментарий отрицательный.