-1. Авторы и материалы

Катя Герасименко, Леша Виняр.

Код – hw_RAKE.py. Там три части, при запуске выполняются все и результаты выводятся на экран. В отчете все результаты также приведены.

0. Текст

Мы взяли текст про биткоин – объяснение разных проблем и вопросов, связанных с ним. Файл Bitcoin.txt. Текст – это часть статьи из BBC (http://www.bbc.com/news/business-42150512)

1. Ручное выделение ключевых слов

	Катя	Леша
Bitcoin	+	+
crypto-currency	+	+
currency	+	-
buy	+	-
anonymity	+	+
business	+	-
value	+	-
user	+	-
growing	+	-
digital	+	-
alternative currency	-	+
mining	-	+
economic bubble	-	+
speculation	-	+
buy Bitcoin	-	+
Bitcoin wallet(s)	-	+
digital currencies	-	+
900% rise	-	+
investment	-	+
crypto-asset	-	+

Из 20 слов, выделенных нами, только три совпали (15%, выделены голубым). Согласие аннотаторов такое низкое из-за разных принципов аннотирования. Катя выделяла ключевые слова относительно конкретного текста и с учетом их встречаемости в этом тексте, а Леша выделял более тематически важные ключевые слова, не принимая во внимание их частоту в тексте. Зеленым выделены слова из ЗС.

2. RAKE

Топ-15 ключевых слов, выделенных базовым RAKE:

- privately-run bank accounts - 16.0
- frenzied investors paying - 9.0
- make illegal purchases - 9.0
- largely exists online - 9.0
- dutch tulip bulbs - 9.0
- classic economic bubble - 9.0
- chicago mercantile exchange - 9.0
- amid volatile trade - 9.0
- judge business school - 9.0
- large financial institutions - 8.5
- attracted people wanting - 8.25
- rules underpinning bitcoin - 7.875

- bitcoin owners don - 7.875
- bitcoin wallets store - 7.875
- dr garrick hileman - 7.666666666666668

Совпадение – формально нулевое, однако есть economic bubble и bitcoin wallets. То есть 2 / 15 (13%) в лучшем случае. Но в целом эти слова не очень ключевые, потому что большинство из них слишком специфичны и не охватывают тематику целиком. Так, мы считаем совсем ошибками следующие выделенные ключевые слова:

- dutch tulip bulbs
- chicago mercantile exchange
- judge business school
- dr garrick hileman

Хорошим ключевым словом, на наш взгляд, является largely exists online. Остальные, хоть и относятся к теме и «ошибками» их не назвать, отражают совсем конкретные детали темы, иногда косвенные (так, large financial institutions относятся к теме биткоина, но довольно косвенно), в то время как словосочетания меньшей длины могли бы дать более общие ключевые слова.

3. Улучшения RAKE

1) Лемматизировать слова
2) Поставить минимальный порог встречаемости – допустим, 3.

Мы сделали лемматизацию с помощью WordNetLemmatizer. Лемматизация не привела bitcoins к bitcoin, но в целом выглядит нормально.
Минимальный порог 3 сработал неадекватно и дал слова bitcoin, user, bitcoins, year и загадочное ``.
Сравним минимальные пороги 1 и 2.

1	2
privately-run bank account - 15.0 classic economic bubble - 9.0 registry link real - 9.0 largely exist online - 9.0 amid volatile trade - 9.0 chicago mercantile exchange - 9.0 dutch tulip bulb - 9.0 make illegal purchase - 8.5 large financial institution - 8.5 judge business school - 8.0 rule underpin bitcoin - 7.8125 dr garrick hileman - 7.666666666666668 frenzied investor pay - 7.666666666666667 bitcoin wallet store - 6.8125 digital currency bitcoin - 5.8125	dr hileman - 4.666666666666667 crypto-currency - 4.0 buy thing - 3.857142857142857 buy - 1.8571428571428572 bitcoin - 1.8125 pay - 1.6666666666666667 address - 1.3333333333333333 world - 1.3333333333333333 user - 1.25 bitcoins - 1.25 day - 1.0 year - 1.0 cambridge - 1.0 create - 1.0 space - 1.0

В минимальном пороге 1 (то есть по сравнению с базовым прибавилась только лемматизация) добавилось одно хорошее ключевое слово - digital currency bitcoin, но добавилось и странное - registry link real (registry linking real [names to addresses]).
В минимальном пороге 2 теперь одиночные слова и несколько биграмм. Самое топовое ключевое слово – имя и не годится, 5 последних ключевых слов тоже не подходят (day, year, create, space слишком общие, чтобы быть ключевыми, Cambridge название и не про биткоины). Остальные вполне в тему, с ЗС 4 пересечения (если считать пересечением и bitcoin, и bitcoins), то есть 27%, и еще есть user, который был в списке у Кати. Но в целом результат тоже не очень хороший – теперь в другую сторону. Хотелось бы что-то среднее между результатами 1 и 2.

В итоге, на этом тексте RAKE показал не самые лучшие результаты, хотя улучшенный алгоритм работает немного получше, чем базовый. Возможно, сам текст не очень удачный и на самом деле не совсем на одну тему.

4. RAKE на русском

Мы взяли похожую статью про биткоин (более техническую, впрочем), с BBC на русском – http://www.bbc.com/russian/business/2016/01/160118_bitcoin_developer_abandoning_project. Файл Bitcoin_rus.txt.

Получили следующий список:

- интервью би-би-си - 16.0
- указанием биткоин-адресов отправителей - 14.222222222222221
- the money project - 9.0
- мере распространения биткоинов - 9.0
- своем блоге текст - 9.0
- андресен является вторым - 9.0
- включая соединенные штаты - 9.0
- увеличить емкость сети - 9.0
- пользователи системы имеют - 9.0
- участия третьей стороны - 9.0
- вызвал всеобщее удивление - 9.0
- порождает ожесточенные споры - 9.0
- соучредитель стартапа blockchain - 9.0
- идее записи данных - 9.0
- существует огромный интерес - 9.0

Из 15 слов ключевыми могут послужить три, остальные слишком частные, детальные и / или косвенно относятся к теме. Как улучшить – русскому больше нужна лемматизация, чем английскому + посмотреть на биграммы (макс. длина 2) или увеличить минимальную частоту. Еще можно кастомизировать список стоп-слов. Так, в них можно попробовать включить слова «являться», «существовать», «иметь» или не включать какие-то слова.