1)

Код улучшенного RAKE-а в файле rake.py

Текст:

*Twitter is an online news and social networking service where users post and interact with messages, known as "tweets." These messages were originally restricted to 140 characters, but on November 7, 2017, the limit was doubled to 280 characters for all languages except Japanese, Korean, and Chinese. Registered users can post tweets, but those who are unregistered can only read them. Users access Twitter through its website interface, Short Message Service (SMS) or mobile device application software ("app"). Twitter, Inc. is based in San Francisco, California, and has more than 25 offices around the world.*

*Twitter was created in March 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams and launched in July of that year. The service rapidly gained worldwide popularity. In 2012, more than 100 million users posted 340 million tweets a day, and the service handled an average of 1.6 billion search queries per day. In 2013, it was one of the ten most-visited websites and has been described as "the SMS of the Internet". As of 2016, Twitter had more than 319 million monthly active users. On the day of the 2016 U.S. presidential election, Twitter proved to be the largest source of breaking news, with 40 million election-related tweets sent by 10 p.m. (Eastern Time) that day.*

*The company experienced rapid initial growth. It had 400,000 tweets posted per quarter in 2007. This grew to 100 million tweets posted per quarter in 2008. In February 2010, Twitter users were sending 50 million tweets per day. By March 2010, the company recorded over 70,000 registered applications. As of June 2010, about 65 million tweets were posted each day, equaling about 750 tweets sent each second, according to Twitter. As of March 2011, that was about 140 million tweets posted daily. As noted on Compete.com, Twitter moved up to the third-highest-ranking social networking site in January 2009 from its previous rank of twenty-second.*

*Twitter's usage spikes during prominent events. For example, a record was set during the 2010 FIFA World Cup when fans wrote 2,940 tweets per second in the thirty-second period after Japan scored against Cameroon on June 14. The record was broken again when 3,085 tweets per second were posted after the Los Angeles Lakers' victory in the 2010 NBA Finals on June 17, and then again at the close of Japan's victory over Denmark in the World Cup when users published 3,283 tweets per second. The record was set again during the 2011 FIFA Women's World Cup Final between Japan and the United States, when 7,196 tweets per second were published. When American singer Michael Jackson died on June 25, 2009, Twitter servers crashed after users were updating their status to include the words "Michael Jackson" at a rate of 100,000 tweets per hour. The current record as of August 3, 2013 was set in Japan, with 143,199 tweets per second during a television screening of the movie Castle in the Sky (beating the previous record of 33,388, also set by Japan for the television screening of the same movie).*

*Twitter acquired application developer Atebits on April 11, 2010. Atebits had developed the Apple Design Award-winning Twitter client Tweetie for the Mac and iPhone. The application, now called "Twitter" and distributed free of charge, is the official Twitter client for the iPhone, iPad and Mac.*

*From September through October 2010, the company began rolling out "New Twitter", an entirely revamped edition of twitter.com. Changes included the ability to see pictures and videos without leaving Twitter itself by clicking on individual tweets which contain links to images and clips from a variety of supported websites including YouTube and Flickr, and a complete overhaul of the interface, which shifted links such as '@mentions' and 'Retweets' above the Twitter stream, while 'Messages' and 'Log Out' became accessible via a black bar at the very top of twitter.com. As of November 1, 2010, the company confirmed that the "New Twitter experience" had been rolled out to all users.*

2)

Список Залины (AZ):

twitter

social networking

interact with messages

post tweets

application software

worldwide popularity

100 million users

tweets

search queries

most-visited

active users

breaking news

election-related

growth

per quarter

million tweets

each day

site

prominent events

record

per second

set again

published

servers crashed

updating

words

developer

free of charge

revamped edition

changes

links

pictures

videos

clicking

interface

links

black bar

new twitter experience

japan

Список Даши (ZD):

TwitterSocial networkingUsersPost tweetsCharactersDevice application softwarePopularitySearch queriesMost-visited websitesElectionSourceBreaking newsGrowthPer daySiteProminent eventsWorld cupRecordPer secondJapanVictoryServers crashedUpdatingStatusTelevision screeningDeveloperClientFree of chargeRevamped editionChangesPictures VideosClicking LinksSupported websitesInterfaceBlack barNew Twitter experienceМаксимальная длина тематической единицы = 3 слова

Элементов в ручном списке: 39 (AZ), 38 (ZD)

Элементов в списке Rake: 178; в т.ч. длиной не более трех слов: 162

Совпадения: az & rake = 18; zd & rake = 0; az & zd = 24

Список участника AZ состоит из 39 слов, список участника DZ – из 38. Общими в двух списках оказались 24 слова, таким образом, процент пересечения двух списков мы можем посчитать так: (AZ ∩ DZ)\*2/( AZ + DZ), а именно 48/77=0,623… То есть, процент совпадения двух списков составил более 60%. При этом стоит отметить, что также были выделены ключевые фразы, которые отличались только одним словом: device application software/application software, users/ 100 million users, popularity/worldwide popularity, most-visited websites/most-visited, election/election-related, per day/each day/per second/per quarter. Это означает, что алгоритмы выборы ключевых слов и фраз у участников очень похожи, а данные слова и фразы можно было бы также отнести к списку пересечений, и тогда бы процент совпадения был бы еще выше.

3)

Top-15 original RAKE:

apple design award

official twitter client

twitter servers crashed

users access twitter

65 million tweets

users post

registered users

twitter users

online news

originally restricted

jack dorsey

noah glass

biz stone

evan williams

breaking news

Ошибочно выделенные ключевые слова:

los angeles lakers

2011 fifa women

san francisco

000 registered applications

fans wrote 2

000 tweets posted

users published 3

movie castle

japan scored

750 tweets

940 tweets

085 tweets

283 tweets

196 tweets

000 tweets

199 tweets

july

year

included

ability

day

ten

cameroon

sky

september

@mentions

Rolled

Как можно заметить, значительная часть неверно выделенных ключевых слов связана с некорректным выделением чисел, в которых в исходном тексте тысячи отделяются запятой и потому разбиваются RAKE-ом на два токена. Также неточности вызваны обрезанием цельного названия (например, 2011 FIFA Women's World Cup Final), и с занесением одиночных неспецифичных слов (day, September).

В качестве улучшения я бы предложила попробовать выделять ключевые слова в лемматизированном тексте, потому что это помогло бы избежать расхождений вроде того, что человек бы выделил какое-то слово, несколько раз повторяющееся в тексте, один раз и записал его в начальной форме, а RAKE может сделать это несколько раз для разных форм (например, include, included).

4)

Top-15 new RAKE:

social networking service

short message service

apple design award

official twitter client

twitter servers crashed

users access twitter

65 million tweets

users post

registered users

twitter users

online news

originally restricted

jack dorsey

noah glass

biz stone

Таким образом если судить по top-15 мы улучшили RAKE на 13%, при чем новым выделившимся тематическим выражениям RAKE приписал самый высокий вес.