Извор: https://www.aclweb.org/anthology/L18-1506.pdf

ТЕРМИНИ:

The Recall-Oriented Understudy for Gisting Evaluation (Rouge)-score -> скаларна вредност во ранг од [0,1]. ја прикажува сличноста помеѓу еден и еден или повеќе токенизирани документи. Го оценува квалитетот на моделите за сумирање.

ВОВЕД

Во воведот се објаснува за претходни истражувања во врска со текст сумаризацијата. Проблемот настанувал поради мали податочни множества за проценка кој е подобар модел за сумаризација. Пример за твитовите, авторите морале сами да создаваат податочни множества. Иако имало множество од твитови со преку 230 000, ипак имало проблем со референците до нив што влијаеле врз проценките на моделот.

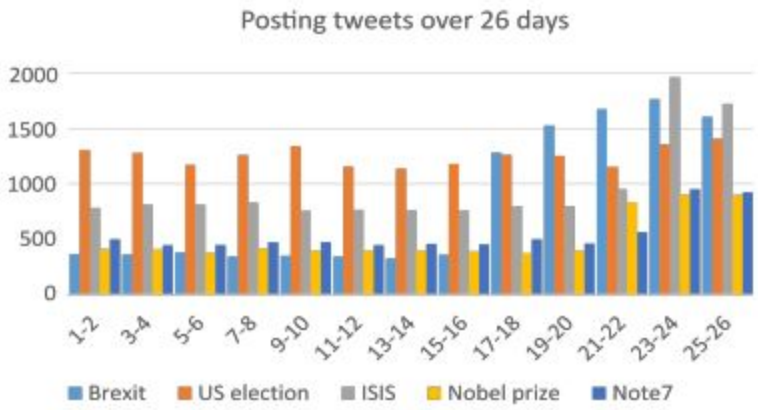
Овој научен труд уствари претставува текст сумаризација, прикажувајќи ново множество на твитови кои обработуваат 6 теми од социјален аспект. За да покажат дека овој dataset е уствари потенцијално употреблив, користат хибриден TF-IDF за извлекување на нови твитови, за да се докаже дека има компетитивен ROUGE скор.

СУМАРИЗАЦИЈА:

Се дели на два чекора: data creation и моделот вклучувајќи и tweet scoring, selection

Data Creation

Значи за да се направи под. М-во потребно е да листа од области кои треба задоволуваат некои услови. Како на пример, да бидат преземени од различни извори, не се постари од 30 дена и мора да бидат некои популарни (импресивни) да речеме..Потоа за секоја област (topic) се доделува листа со клучни зборови, поединечно.



Event	#tweets	#hashtags	Keywords
Brexit	10,978	9,705	brexit, #brexitshambles, #Brexiters, #BrexitCentral, England Europe exit.
US election	17,714	8,566	Donald Trump, Trump, Hilary Clinton, Clinton #debate, election.
ISIS	13,047	9488	ISIS, IS Syria, IS Mosul, IS Iraq, ISIS Aleppo, ISIS US, ISIS Rusia.
SS Note 7	7,362	7,465	Galaxy Note 7, #note7, #GalaxyNote7, thegalaxynote7, #SamsungGalaxyNote7.
Nobel prize	6,812	2,780	Nobel prize 2016, Nobel peace, Nobel chemistry, Nobel economy, Nobel physics.
SpaceX	4,982	2,417	Facebook SpaceX, SpaceX Explosion, Falcon 9 exploded, Falcon 9 explosion.

Ова е распределба на твитови. Покажува статистика за користење на клучните зборови.

МОЖЕ ДА БИДАТ ИСКОРИСТЕНИ ЗА ДА СЕ ПРОЦЕНИ КВАЛИТЕТОТ НА ТВИТОВИТЕ, МЕРЕЈЌИ КОЛКУ Е БИТНА ИНФОРМАЦИЈАТА КОЈА Е ГЕНЕРИРАНА ОД КОРИСНИЦИТЕ НА ТВИТОВИТЕ.

ПОДАТОЧНА СИГМЕНТАЦИЈА:

Собраните твитови се делат во кластери. Идејата е таква бидејќи дури и бројот на твитови на ден да е мал, директно извлекување на подмножество од овие твитови можеби ќе ги елиминира останатите кои се битни. Значи… со кластерирањето, целта е да твитовите бидат што порепрезентивни.
Е сега за да биде се тоа во реално време од аспект на твит сумаризацијата, се користи алгоритам **Affinity Propagation , алгоритам за кластерирање.**
AP алгоритамот, го идентификува подмножеството од data points како темплејтови и формира кластери така што ги доделува останатите data points во еден од тие темплејтови (exemplars). После кластерирањето и елиминирање на тие денови со малку твитови, се формирале 6 множества кои кореспондираат на 6те настани во тие 26 дена и секој ден вклучува м-во од кластери, кои на некој начин се подобласти.

Standard reference creation

Ова се користи за еволуација на моделот. Се користат 3 методи
1.Luhn кој е хевристички метод за извлекување на тие податоци.
2. LexRank е графички-базиран модел, кој гради граф на сличност и ги селектира оние кои се засноваат на нивниот сопствен вектор (ова е од линеарна алгебра и сега не е толку битно да го знаеш ти).
3. DSDR-non се базира на ненегативна линеарна податочна реконструкција.

Овие извлечени твитови од овие три методи формираат множества кои се кандидати.

Method	ROUGE-1	ROUGE-2	ROUGE-SU
Luhn	0.556	0.502	0.250
LexRank	0.581	0.516	0.273
DSDR-non	0.419	0.310	0.145

Според ова LexRank е најдобар метод, бидејќи има најголема ROUGE вредност

Обзервација на податоците: