Xiao's Solution

Abstract

Given this problem, the first idea came to my mind is to use a supervised method to predict the topics. However, when having this unlabeled dataset, it is usually not a good idea to manually label the instance. After some research on similar problems focusing on clustering and statistical models, finally the TF-IDF algorithm is selected to generate the result. In the following sections I'll discuss my approach in detail.

Tool Used

Java 1.8, Apache Lucene Analysis, LibreOfiice Vanilla, Python

Steps

- Preprocess Data, deal with empty space and null value, parse all the words in article;
- Keyword candidate selection; Because not all words in the file is valuable(e.g. 'a', 'at'), it is critical to find a list of potential keywords which may help improving accuracy in future. In this approach, keywords are mainly filtered by this four schemes:
  - Remove numbers and special characters;
  - Remove stop words using a stop list;
  - Stem the words to make similar words the same(e.g. 'study' and 'studies');
  - Evaluate if a phrase is a candidate(e.g. 'weight loss');
- Use bag of words to encode the instance. Different weights are applied for title and body; Calculate the TF-IDF value and list the top k-words;
- Decode the stemmed k-words, present the original words as topics;

Implementation Detail

1. Select the candidates and generate a stemmed article as showing below. In the figure, number and special characters are removed. Words are stemmed into weird words;

why good diet nutrut import	dietari factor singl most signific risk factor disabl prematur death despit wealth knowledg link food l
lower total fat intak caus weight loss	studi research found lower total fat intak lead sustain reduct bodi weight adult children durat six month eight
supplement food substitu benefit	supplement aren intend food substitut older adult other mai benefit specif supplement
yoga ha mani mental physic benefit	yoga provid retreat chaotic busi live yoga provid mani other mental physic benefit benefit extend eat cook h
make meal whole grain full fiber	good daili nutrit try make meal whole grain low sodium low sugar fiber rich low calori
kraft singl been label kid eat right	kraft singl first food receiv new kid eat right label stamp approv design academi nutrit dietet help famili make
lack vitamin d ha been associ mani cancer heart disea	nearli half world popul suffer vitamin d defici lack vitamin d ha been associ cancer heart diseas diabet multi
coffe part healthi lifestyl	coffe doesn deserv it dark histori moder consumpt mai incorpor healthi lifestyl bean like seed insid coffe pla
peopl eat vegetarian diet lower risk colorect cancer	research analyz dietari habit peopl at vegetarian diet had lower risk colorect cancer among at vegetarian di
exercis weight loss import rest energi expenditur	regular exercis boost rest energi expenditur rate which burn calori workout rest benefit burn extra energi wh
fibr consumpt significantli aid weight loss	increas fiber consumpt significantli aid weight loss lower blood pressur enhanc insulin resist fast insulin
tip start own organ garden	grow own fruit veget almost us probabl grade school plant seed cup dirt water watch grow what distinguish
understand food label make easier make right food ch	know what eat understand food label make easier consum make quick inform food choic contribut healthi di
refriger thermomet cold fact food safeti	on most effect tool protect yourself famili food born ill refriger room temperatur number bacteria caus food b
low carb diet effect low fat diet	mediterranean low carbohydr diet mai effect altern low fat diet low carbohydr diet favor effect lipid mediterra
long term weight loss mainten	approxim overweight individu success lose least initi bodi weight maintain loss least yr
milk chees good	research ha shown longer necessari milk dairi product diet better consum healthi diet grain fruit veget legun

2. Bag of Word model. Once all words are stemmed(because we don't want 'study' and 'studies' being different words), the occurrence of each word is counted and the whole data was encoded to expected input; To be mentioned, because title and body are different attributes, different weights are given to the word in attribute. Encoded data is shown below;

5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

3. TF-IDF algorithm is used to determined the importance of a word in an article. While usually used as a weight algorithm, it can also be used to classify hot topics in this problem. Selected topics are shown on the left side; And decoded data is on the right side;

nutrut	why	import
fat intak	total	lower
supplement	substitu	benefit
yoga	mental	benefit
whole grain	fiber	meal
kraft	label	singl
vitamin d	cold	lack
coffe	lifestyl	healthi
vegetarian	colorect	lower risk
expenditur	rest	energi
significantli	fibr	consumpt
organ	garden	start
make	food	easier
refriger	thermomet	food
diet	effect	low fat

nutrution	why	important
fat intake	total	lowering
supplements	substitues	benefits
yoga	mental	benefits
whole grain	fiber	meals
kraft	labelled	singles
vitamin d	cold	lack
coffee	lifestyle	healthy
vegetarian	colorectal	lower risk
expenditure	resting	energy
significantly	fibre	consumption
organic	garden	start
makes	food	easier
refrigerator	thermometers	food
diets	effective	low fat

4. However, problem still exists when decoding because it is hard to turn the words back; The un-decoded words shown below are not important though, a good solution is to simply include them in the stop list when selecting candidates;

banana	didn	thing
younger	heart attack	women
obes	five	overweight
sugar	isn	back

bananas		things
younger	heart attacks	women
obese	five	overweight
sugar		back

Un-implemented part and ideas

1. Use a smoothing algorithm to add weight to words in the sentence according to the position of the words. Because research showed words in the begin or at the end of a sentence will have more probability being a keyword;

2. Cluster the data and see the result;