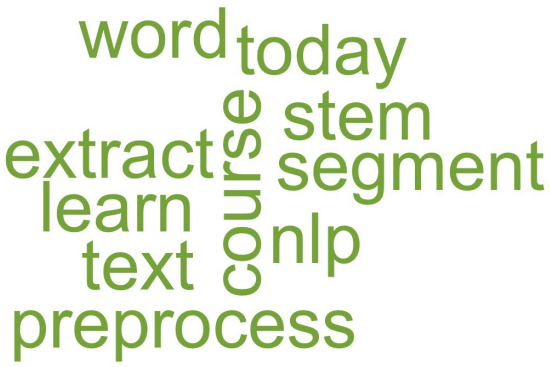
1. TF scores、word cloud image:

The word cloud images of text0, text1, text2, text3, text4, text5, text8 and text9 are similar due to all words have the same frequency. So, here only report word cloud images of text0, text6 and text7.

text0

today	go	nlp	course	learn	segment	word	extract	stem	preprocess	text
0.09091	0.09091	0.09091	0.09091	0.09091	0.09091	0.09091	0.09091	0.09091	0.09091	0.09091



text1

today	go	school	want	participant	text	analyse	course	sound	nice	due	segment	word
0.07692	0.07692	0.07692	0.07692	0.07692	0.07692	0.07692	0.07692	0.07692	0.07692	0.07692	0.07692	0.07692

text2

yesterday	go	data	science	course	learnt	preprocess	text	happy	due	teacher	nice
0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333

text3

today	nlp	course	interesting	learnt	lot	knowledge	useful	career	happy
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

text4

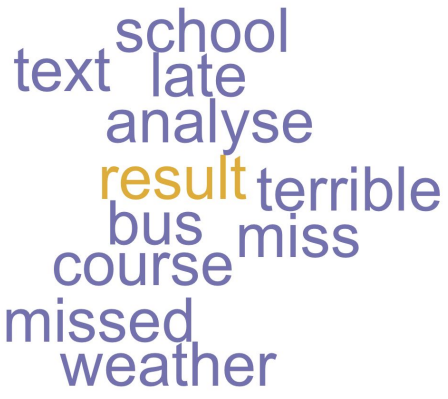
today	go	school	want	participant	data	mining	course	sound	nice	due	secret
0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333

text5

machine	learn	course	also	nice	teacher	pleased	professional	teach	u	useful	technology
0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333

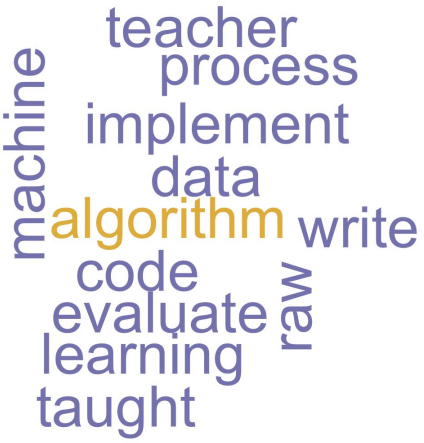
text6

weather	terrible	result	missed	bus	late	school	miss	text	analyse	course
0.08333	0.08333	0.16667	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333



text7

teacher	taught	u	process	raw	data	write	code	implement	algorithm	evaluate	machine	learning
0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.14286	0.07143	0.07143	0.07143



text8

today	weather	nice	go	school	absence	machine	learn	course	climb	mountain	picnic
0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333	0.08333

text9

today	go	gym	sport	finish	homework	assign	machine	data	mining	course	feel	well	exercise
0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143	0.07143

TF-IDF scores

Text0

today	go	nlp	course	learn	segment	word	extract	stem	preprocess	text
0.03242	0.03242	0.10945	0.0	0.0833	0.10945	0.10945	0.14631	0.14631	0.10945	0.06301

Text1

today	go	school	want	participant	text	analyse	course	sound	nice	due	segment	word
0.02744	0.02744	0.05332	0.09261	0.09261	0.05332	0.09261	0.0	0.09261	0.03929	0.07048	0.09261	0.09261

Text2

yesterday	go	data	science	course	learnt	preprocess	text	happy	due	teacher	nice
0.13412	0.02972	0.05776	0.13412	0.0	0.10033	0.10033	0.05776	0.10033	0.07636	0.07636	0.04257

Text3

today	nlp	course	interesting	learnt	lot	knowledge	useful	career	happy
0.03567	0.1204	0.0	0.16094	0.1204	0.16094	0.16094	0.1204	0.16094	0.1204

Text4

today	go	school	want	participant	data	mining	course	sound	nice	due	secret
0.02972	0.02972	0.05776	0.10033	0.10033	0.05776	0.10033	0.0	0.10033	0.04257	0.07636	0.13412

Text5

machine	learn	course	also	nice	teacher	pleased	professional	teach	u	useful	technology
0.05776	0.07636	0.0	0.13412	0.04257	0.07636	0.13412	0.13412	0.13412	0.10033	0.10033	0.13412

Text6

weather	terrible	result	missed	bus	late	school	miss	text	analyse	course
0.10033	0.13412	0.26824	0.13412	0.13412	0.13412	0.05776	0.13412	0.05776	0.10033	0.0

Text7

teacher	taught	u	process	raw	data	write	code	implement	algorithm	evaluate	machine	learning
0.06545	0.11496	0.086	0.11496	0.11496	0.04951	0.11496	0.11496	0.11496	0.22992	0.11496	0.04951	0.11496

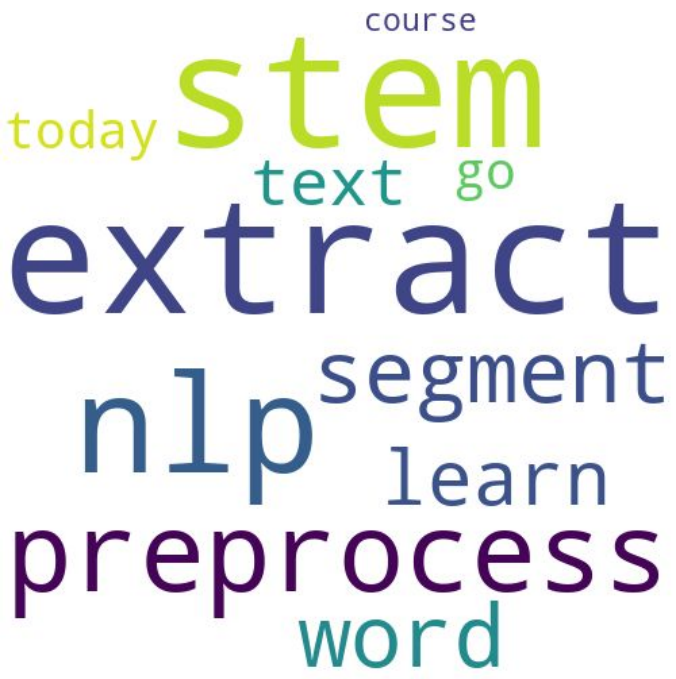
Text8

today	weather	nice	go	school	absence	machine	learn	course	climb	mountain	picnic
0.02972	0.10033	0.04257	0.02972	0.05776	0.13412	0.05776	0.07636	0.0	0.13412	0.13412	0.13412

Text9

today	go	gym	sport	finish	homework	assign	machine	data	mining	course	feel	well	exercise
0.02548	0.02548	0.11496	0.11496	0.11496	0.11496	0.11496	0.04951	0.04951	0.086	0.0	0.11496	0.11496	0.11496

Word cloud after TF-IDF:



2. Top-10 pairs PMI scores:

```
[(('bus', 'late'), 6.930737337562887), (('climb', 'mountain'), 6.930737337562887), (('code', 'implement'), 6.930737337562887), (('extract', 'stem'), 6.930737337562887), (('feel', 'well'), 6.930737337562887), (('finish', 'homework'), 6.930737337562887), (('gym', 'sport'), 6.930737337562887), (('homework', 'assign'), 6.930737337562887), (('lot', 'knowledge'), 6.930737337562887), (('missed', 'bus'), 6.930737337562887)]
```

I don't think this result makes sense. Because when the text is large, two-pairs are very numerous and will be common, and there are many pairs that only appear together once. This result is meaningless. Maybe adjusting to four-pairs might be better.

```
[(('gym', 'sport', 'finish', 'homework'), 20.79221201268866), (('sport', 'finish', 'homework', 'assign'), 20.79221201268866), (('code', 'implement', 'algorithm', 'evaluate'), 19.79221201268866), (('interesting', 'learnt', 'lot', 'knowledge'), 19.79221201268866), (('lot', 'knowledge', 'useful', 'career'), 19.79221201268866), (('pleased', 'professional', 'teach', 'u'), 19.79221201268866), (('result', 'missed', 'bus', 'late'), 19.79221201268866), (('taught', 'u', 'process', 'raw'), 19.79221201268866), (('terrible', 'result', 'missed', 'bus'), 19.79221201268866), (('write', 'code', 'implement', 'algorithm'), 19.79221201268866)]
```

3.

The program code:

```
def ent(data):
   p_data = data.value_counts()
   entropy = scipy.stats.entropy(p_data)
   return entropy
```

The Twitter data for advert:

```
spam_set.append('Let the Cameras In - Stop secret heath care negotiations')
spam_set.append('Stop secret heath care negotiations. Sign the petition to let the Cameras In!')
spam_set.append('Let the Cameras In - Stop secret heath care negotiations')
```

```
spam_set.append('Sign the petition - Let the Cameras In!')
spam_set.append('What are Speaker Pelosi,Sen. Reid and Pres Obama hiding?')
spam_set.append('Sign the petition - Let the Cameras In!')
spam_set.append('Sign the petition - Let the Cameras In!')
spam_set.append('No Secret health care negotiations!')
spam_set.append('No more secret health care negotiations! Tranparency now.')
spam_set.append('Let the Cameras In - Stop the secret health care negotiations')
```

Entropy values for spam-set= 1.8343719702816235

The random data:

```
random_set=[]
random_set.append('This is a nice diagram by Zhengyan Zhang and @BakserWang that shows how many recent pretrained language models are connected. The GitHub repo contains a full list of relevant papers')
random_set.append('Working in partnership with @hotpress to curate a list of extraordinary tracks to inspire you in  the lead up to World Mental Health Day. Listen now on Spotify.')
random_set.append('A Vardy household source tells me that Colleen\'s message could actually put Rebekah in a  stronger position')
random_set.append('Imagining the Vardy household right now, Rebekah shrieking blue murder, Jamie\'s on his ninth  Red Bull of the morning, kids are crying')
random_set.append('Leo Varadkar has reneged on a secret deal with Boris Johnson to open the way to a Brexit compromise, a senior Downing Street source claimed yesterday')
random_set.append('It was touch and go but Wales have won it! Full-time Wales 29-17 Fiji')
random_set.append('It\'s up to you today to start making healthy choices. Not choices that are just healthy for your body, but healthy for your mind.')
random_set.append('I\'ve already perfect the art of fake smiling.')
random_set.append('VIDEO: Police and demonstrators clash outside Ecuador\'s National Assembly building as protests over a fuel hike introduced by President Lenin Moreno\'s government intensify')
random_set.append('How are wind farms installed in the sea?')
```

Entropy values for random-set= 2.3025850929940455

Entropy values for mixed-set= 2.7616257121977803