# Text summarization using Gensim

Genism was used for giving automatically summarizes the given text, by extracting one or more important sentences from the text. In a similar way, it can also extract keywords. This tutorial will teach you to use this summarization module via some examples. First, we will try a small example, then we will try two larger ones, and then we will review the performance of the summarizer in terms of speed.
This summarizer is based on the "TextRank" algorithm.
As text is pre-processed so that stopwords are removed and the words are stemmed so Gensim's summarization works for English for now and these processes are language-dependent.

Why we need to introduce PageRank before TextRank? Because the idea of TextRank comes from PageRank and using similar algorithm (graph concept) to calculate the importance.

**Difference:**

- TextRank graph is undirected. Meaning that all edge are bidirectional
- The weight of edge is difference while it is 1 in PageRank. There are different way to calculate such as BM25, TF-IDF.

### Why genism ?

- Gensim is a well-optimized library for topic modeling and document similarity analysis..Even so, it's a valuable tool to add to your repertoire. Its topic modeling algorithms, such as its Latent Dirichlet Allocation (LDA) implementation, are best-in-class. In addition, it's robust, efficient, and scalable.

- Plus, the sub-field semantics analysis (or topic modeling), is one of the most exciting areas of modern natural language processing.

- The gensim implementation is based on the popular TextRank algorithm.

## Source code

gensim provides a simple API to calculate TextRank by using BM25 (Best Match 25).

### Step 1: Environment Setup

pip install gensim==3.8.0

```
In [1]: import gensim
        print('gensim Version: %s' % (gensim.__version__))

        gensim Version: 3.8.0
```

## Step 2: Import data

In [2]:
```
# sample data

content = "Microsoft held talks in the past few weeks " + \
    "to acquire software developer platform GitHub, Business " + \
    "Insider reports. One person familiar with the discussions " + \
    "between the companies told CNBC that they had been " + \
    "considering a joint marketing partnership valued around " + \
    "$35 million, and that those discussions had progressed to " + \
    "a possible investment or outright acquisition. It is " + \
    "unclear whether talks are still ongoing, but this " + \
    "person said that GitHub's price for a full acquisition " + \
    "was more than Microsoft currently wanted to pay. GitHub " + \
    "was last valued at $2 billion in its last funding round " + \
    "2015, but the price tag for an acquisition could be $5 " + \
    "billion or more, based on a price that was floated " + \
    "last year. GitHub's tools have become essential to " + \
    "software developers, who use it to store code, " + \
    "keep track of updates and discuss issues. The privately " + \
    "held company has more than 23 million individual users in " + \
    "more than 1.5 million organizations. It was on track to " + \
    "book more than $200 million in subscription revenue, " + \
    "including more than $110 million from companies using its " + \
    "enterprise product, GitHub told CNBC last fall.Microsoft " + \
    "has reportedly flirted with buying GitHub in the past, " + \
    "including in 2016, although GitHub denied those " + \
    "reports. A partnership would give Microsoft another " + \
    "connection point to the developers it needs to court to " + \
    "build applications on its various platforms, including " + \
    "the Azure cloud. Microsoft could also use data from " + \
    "GitHub to improve its artificial intelligence " + \
    "producs. The talks come amid GitHub's struggle to " + \
    "replace CEO and founder Chris Wanstrath, who stepped " + \
    "down 10 months ago. Business Insider reported that " + \
    "Microsoft exec Nat Friedman -- who previously " + \
    "ran Xamarin, a developer tools start-up that Microsoft " + \
    "acquired in 2016 -- may take that CEO role. Google's " + \
    "senior VP of ads and commerce, Sridhar Ramaswamy, has " + \
    "also been in discussions for the job, says the report. " + \
    "Microsoft declined to comment on the report. " + \
    "GitHub did not immediately return a request for comment."
```

## Step 3 : processing the output on number of words

In [4]:
```
#sumarizing with the no of total words in summary
print('Original Content:')
print(content)
for word_count in [30, 50, 100]:
    summarized_content = gensim.summarization.summarize(content, word_count=word_count)
    print()
    print('---> Summarized Content (Word Count is %d):' % word_count)
    print(summarized_content)
```

```
Original Content:
Microsoft held talks in the past few weeks to acquire software developer platform GitHub, Business Insider reports. One person familiar with the discussions between the companies told CNBC that they ha
d been considering a joint marketing partnership valued around $35 million, and that those discussions had progressed to a possible investment or outright acquisition. It is unclear whether talks are s
till ongoing, but this person said that GitHub's price for a full acquisition was more than Microsoft currently wanted to pay. GitHub was last valued at $2 billion in its last funding round 2015, but t
he price tag for an acquisition could be $5 billion or more, based on a price that was floated last year. GitHub's tools have become essential to software developers, who use it to store code, keep tra
ck of updates and discuss issues. The privately held company has more than 23 million individual users in more than 1.5 million organizations. It was on track to book more than $200 million in subscrip
tion revenue, including more than $110 million from companies using its enterprise product, GitHub told CNBC last fall.Microsoft has reportedly flirted with buying GitHub in the past, including in 201
6, although GitHub denied those reports. A partnership would give Microsoft another connection point to the developers it needs to court to build applications on its various platforms, including the Az
ure cloud. Microsoft could also use data from GitHub to improve its artificial intelligence producs. The talks come amid GitHub's struggle to replace CEO and founder Chris Wanstrath, who stepped down 1
0 months ago. Business Insider reported that Microsoft exec Nat Friedman -- who previously ran Xamarin, a developer tools start-up that Microsoft acquired in 2016 -- may take that CEO role. Google's se
nior VP of ads and commerce, Sridhar Ramaswamy, has also been in discussions for the job, says the report. Microsoft declined to comment on the report. GitHub did not immediately return a request for c
omment.

---> Summarized Content (Word Count is 30):
Microsoft held talks in the past few weeks to acquire software developer platform GitHub, Business Insider reports.

---> Summarized Content (Word Count is 50):
Microsoft held talks in the past few weeks to acquire software developer platform GitHub, Business Insider reports.
One person familiar with the discussions between the companies told CNBC that they had been considering a joint marketing partnership valued around $35 million, and that those discussions had progresse
d to a possible investment or outright acquisition.

---> Summarized Content (Word Count is 100):
Microsoft held talks in the past few weeks to acquire software developer platform GitHub, Business Insider reports.
One person familiar with the discussions between the companies told CNBC that they had been considering a joint marketing partnership valued around $35 million, and that those discussions had progresse
d to a possible investment or outright acquisition.
It was on track to book more than $200 million in subscription revenue, including more than $110 million from companies using its enterprise product, GitHub told CNBC last fall.Microsoft has reportedly
flirted with buying GitHub in the past, including in 2016, although GitHub denied those reports.
```

# other example of the summarization

In [1]:

```
from gensim.summarization.summarizer import summarize
>>> text = '''The desire for more positive experience is itself a negative experience. And, paradoxically,
the acceptance of one's negative experience is itself a positive experience.

The more you pursue feeling better all the time, the less satisfied you become, as pursuing something only reinforces the fact that you lack it in the first place.
Philosopher Alan Watts used to refer to as "The BackWards Law" (further reading: the hedonic treadmill).

Everything worthwhile in life is won through surmounting the associated negative experience.

Pain and loss are inevitable and we should let go of trying to resist them.
The greatest truths in life are usually the most unpleasant to hear.
We suffer for the simple reason that suffering is biologically useful. It is nature's preferred agent for inspiring change.
Don't hope for a life without problems. There's no such thing. Instead, hope for a life full of good problems.
Problems never stop; they merely get exchanged and/or upgraded.
Happiness comes from problems you enjoy having and solving.
Nobody who is actually happy has to stand in front of a mirror and tell himself that he's happy.
Emotions are simply biological signals designed to nudge you in the direction of beneficial change.'''
>>> print(summarize(text))
```

```
Don't hope for a life without problems.
Instead, hope for a life full of good problems.
Emotions are simply biological signals designed to nudge you in the direction of beneficial change.
```

In [ ]: