

SURV703 – Content Analysis

Installing GENSIM, implementing word2vec embedding model, briefly explaining the results

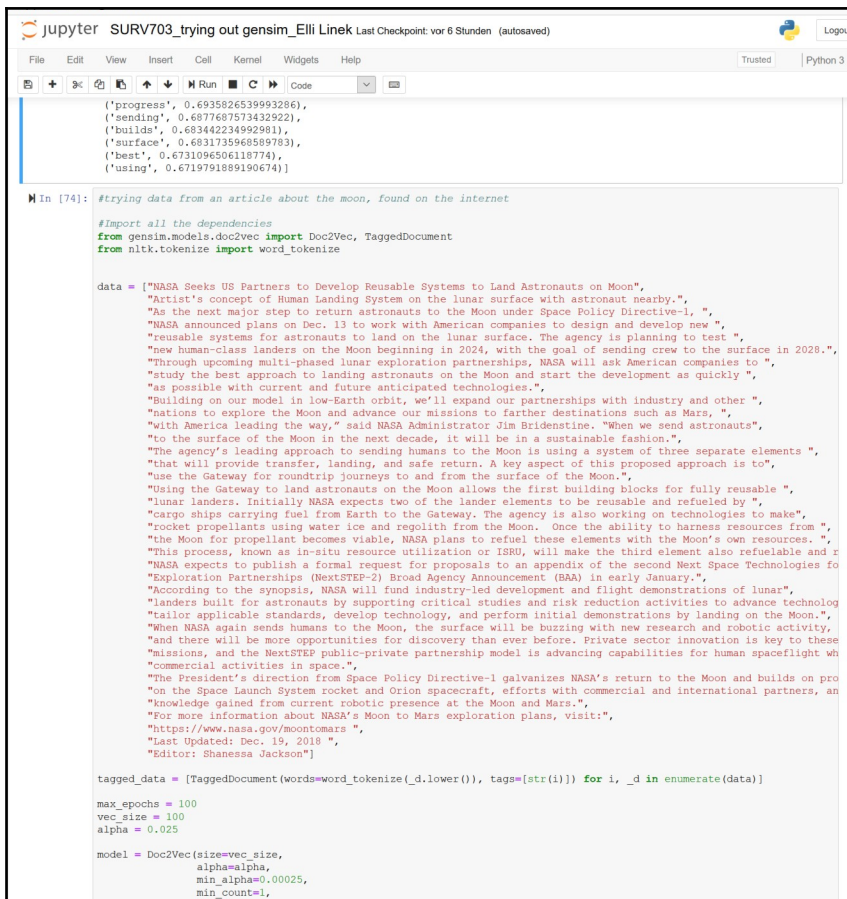
Elisabeth Linek

Task: Using gensim (Links to an external site.), train an LDA topic model (Links to an external site.) or a word2vec embedding model (Links to an external site.) on a corpus of your choice. Interpret the results in a short text of variable length (no more than 500 words).

Starting off with the installation of gensim. The first problem I encountered was, the impossibility to define pdf-files as corpus. So I setted up a short .txt-file, which I was able to load into the program, but encountered later on problems with the set up of the word2vector mode. So is started a completely new trial.

Aiming to test the word embedding possibilities of gensim I changed plans and started with the import of a short text about the moon, that I found on the internet – a short text, not at all complex in order to gain overview and having the possibility to assess the results.

Here is a screenshot out of my jupyter notebook, showing parts of the text, I entered in order to train and test the program:



```
jupyter SURV703_trying out gensim_Elili Linek Last Checkpoint: vor 6 Stunden (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
Run Code

('progress', 0.6935826539993286),
('sending', 0.6877687573432922),
('builds', 0.683442234992981),
('surface', 0.6831735968589793),
('best', 0.6731096506118774),
('using', 0.6719791889190674)]

In [74]: #trying data from an article about the moon, found on the internet

#Import all the dependencies
from gensim.models.doc2vec import Doc2Vec, TaggedDocument
from nltk.tokenize import word_tokenize

data = ["NASA Seeks US Partners to Develop Reusable Systems to Land Astronauts on Moon",
"Artist's concept of Human Landing System on the lunar surface with astronaut nearby.",
"As the next major step to return astronauts to the Moon under Space Policy Directive-1, ",
"NASA announced plans on Dec. 13 to work with American companies to design and develop new ",
"reusable systems for astronauts to land on the lunar surface. The agency is planning to test ",
"new human-class landers on the Moon beginning in 2024, with the goal of sending crew to the surface in 2028.",
"Through upcoming multi-phased lunar exploration partnerships, NASA will ask American companies to ",
"study the best approach to landing astronauts on the Moon and start the development as quickly ",
"as possible with current and future anticipated technologies.",
"Building on our model in low-Earth orbit, we'll expand our partnerships with industry and other ",
"partners to explore the Moon and advance our missions to farther destinations such as Mars, ",
"with America leading the way," said NASA Administrator Jim Bridenstine. "When we send astronauts",
"to the surface of the Moon in the next decade, it will be in a sustainable fashion.",
"The agency's leading approach to sending humans to the Moon is using a system of three separate elements ",
"that will provide transfer, landing, and safe return. A key aspect of this proposed approach is to ",
"use the Gateway for roundtrip journeys to and from the surface of the Moon.",
"Using the Gateway to land astronauts on the Moon allows the first building blocks for fully reusable ",
"lunar landers. Initially NASA expects two of the lander elements to be reusable and refueled by ",
"cargo ships carrying fuel from Earth to the Gateway. The agency is also working on technologies to make",
"rocket propellants using water ice and regolith from the Moon. Once the ability to harness resources from ",
"the Moon for propellant becomes viable, NASA plans to refuel these elements with the Moon's own resources. ",
"This process, known as in-situ resource utilization or ISRU, will make the third element also refuelable and r",
"NASA expects to publish a formal request for proposals to an appendix of the second Next Space Technologies fo",
"Exploration Partnerships (NextSTEP-2) Broad Agency Announcement (BAA) in early January.",
"According to the synopsis, NASA will fund industry-led development and flight demonstrations of lunar",
"landers built for astronauts by supporting critical studies and risk reduction activities to advance technolog",
"tailor applicable standards, develop technology, and perform initial demonstrations by landing on the Moon.",
"When NASA again sends humans to the Moon, the surface will be buzzing with new research and robotic activity",
"and there will be more opportunities for discovery than ever before. Private sector innovation is key to these",
"missions, and the NextSTEP public-private partnership model is advancing capabilities for human spaceflight wh",
"commercial activities in space.",
"The President's direction from Space Policy Directive-1 galvanizes NASA's return to the Moon and builds on pro",
"on the Space Launch System rocket and Orion spacecraft, efforts with commercial and international partners, an",
"knowledge gained from current robotic presence at the Moon and Mars.",
"For more information about NASA's Moon to Mars exploration plans, visit:",
"https://www.nasa.gov/moontomars ",
"Last Updated: Dec. 19, 2018 ",
"Editor: Shanesha Jackson"]

tagged_data = [TaggedDocument(words=word_tokenize(d.lower()), tags=[str(i)]) for i, _d in enumerate(data)]

max_epochs = 100
vec_size = 100
alpha = 0.025

model = Doc2Vec(size=vec_size,
                alpha=alpha,
                min_alpha=0.00025,
                min_count=1,
                da=1)
```

I run 100 iterations, in order to train the model. Based on the then established word to vector model I tested different word-similarities. In a first attempt I asked for similarities regarding the word or term “moon”, and received the following matrix of similar terms:

```
model.train(documents, total_examples=len(documents), epochs=10)
train(...)

In [12]: word = "moon"
model.wv.most_similar (positive = word)

C:\Users\elli\Anaconda3\lib\site-packages\gensim\matutils.py:737: FutureWarning: Conversion of the second argument of
issubdtype from 'int' to 'np.signedinteger' is deprecated. In future, it will be treated as 'np.int32 == np.dtype(int
).type'.
    if np.issubdtype(vec.dtype, np.int):

Out[12]: [('the', 0.84748828411023),
('approach', 0.7347332239151001),
('humans', 0.7176517248153687),
('to', 0.7012702226638794),
('progress', 0.6983550786972046),
('builds', 0.694743275642395),
('sending', 0.6897841095924377),
('best', 0.6805385947227478),
('using', 0.6788151264190674),
('surface', 0.6768609285354614)]
```

In order to have a comparison and for better understanding what gensim does or sets into a relation here I tested another word and selected the term “element” to do so, leading to the following results:

```
In [14]: word2 = "element"
model.wv.most_similar (positive = word2)

C:\Users\elli\Anaconda3\lib\site-packages\gensim\matutils.py:737: FutureWarning: Conversion of the second argument of
issubdtype from 'int' to 'np.signedinteger' is deprecated. In future, it will be treated as 'np.int32 == np.dtype(int
).type'.
    if np.issubdtype(vec.dtype, np.int):

Out[14]: [('third', 0.987095832824707),
('make', 0.9315939545631409),
('refuelable', 0.9237496852874756),
('isru', 0.8931676149368286),
('or', 0.8778015375137329),
('also', 0.8760696053504944),
('utilization', 0.8195888996124268),
('resource', 0.7816526293754578),
('in-situ', 0.7010257840156555),
('working', 0.6892534494400024)]
```

As both results show, I have stop words within the matrix of similar word, that could have been avoided by further pre-processing steps, which I left aside in the run up of the modelling. These stop words are not interpretable. But I received some terms with a content related sense, such as “refueable” - “element” or “third” - “element”. These combinations are very likely to happen, as the accompanied score shows.

Comparable results for moon: Some stop words are included, and most often the prefix “the”-“moon” are accompanied with each other. Not surprisingly, but still, to me as a first contact experience it shows gensim brings up reasonable results.

Concluding, I would say that genism is an astonishing tool, with a lot of possibilities, that I just scratched on but did not at al were digging deeper into. Regarding the results I would have reached more detailed similarities if I would have had done further pre-processing, such as lowercasing, stemming or the removal of small words or stop words. And I assume, that the statistical values would have reached another level if the model would have been trained and tested based on a larger text, which is from a mathematical perspective quite logical.