The Lexicon of the Environment in Social Networks Discourse.

Twitter Data as a Linguistic Reference: Extraction, Cleaning, Usage.

Nikolay Chepurnykh

Tomara Gotkova

ATILF, 08/06/2021

GENERAL INFORMATION ABOUT THE PROJECT

Name of the project: The Lexicon of the Environment and Chemistry-related Terms in Ordinary Discourse. Using Social Networks as Corpora.

Objectives:

Study of core vocabulary of the environment and chemistry-related terms in texts extracted from *Twitter* and *Reddit*:

- Search for potential distortions of scientific vocabulary.
- Search for new senses of scientific vocabulary present in social networks discourse.

TWITTER CORPUS – GENERAL INFORMATION

Extraction technique - keywords list
Dates – **02.09.2020 – 21.09.2020**
Raw collection – **4 709 015 tweets,**
Cleaned collection - **2 860 507 tweets**, **42 155 532 tokens**

created_at	id	text	user.id	original text
Wed Sep 02 10:26:08 +0000 2020	1301104112097001473	It’s time now to build a positive case for a zero-carbon transition to circular economies based on recycling, remanufacturing, reuse, and shared services. and colleagues opinion piece publishes today:	779354632296493056	It’s time now to build a positive case for a zero-carbon transition to circular economies based on recycling, remanufacturing, reuse, and shared services. 🌎♻️ @kbelesova and colleagues @bmj_latest opinion piece publishes today: https://t.co/Dpvwgw18Wi #GreenRecovery #COVID19 https://t.co/jor5fazxWJ
Date of tweet creation	Tweet ID	Tweet text	User’s ID	Tweet text before processing

TWITTER CORPUS – CLEANING STEPS

ACCOMPLISHED:

1. Removing tweets with no keywords
2. Removing full duplicates
3. Removing hashtag symbols; mentions; urls; emojis
4. Removing short tweets (threshold?)

EXAMPLES OF CORPUS USAGE

1. Subcorpus ‘Carbon’

   Senses of carbon present in social networks texts

   Semantic links between carbon and other related terms

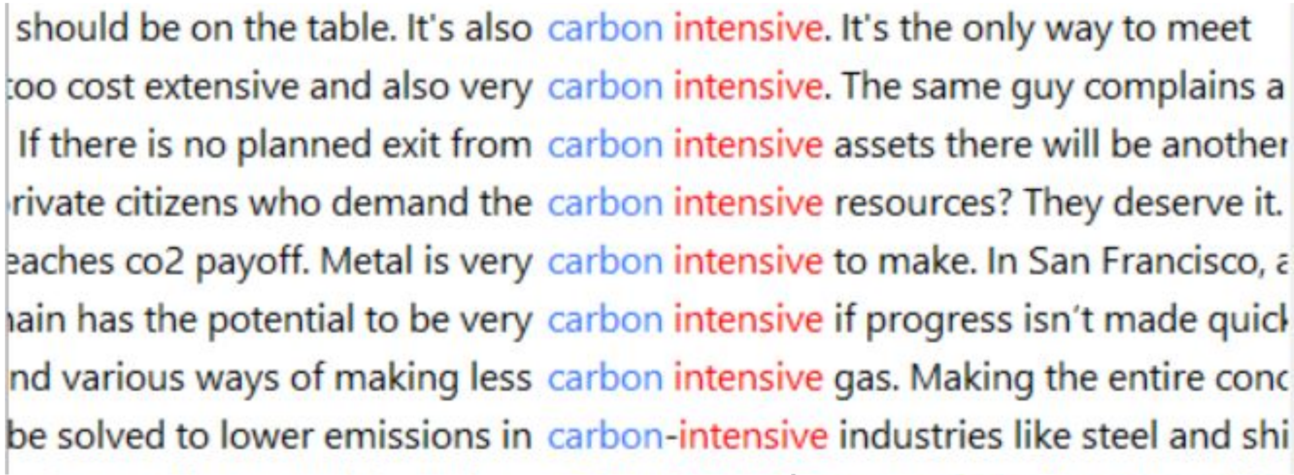
Search query: *carbon* + any(*carbon dioxide, methane, CO2, CH4*)

“**Carbon** in the **methane** then detaches from the **methane** after being in the atmosphere comes back to earth and is reabsorbed by the plants.”

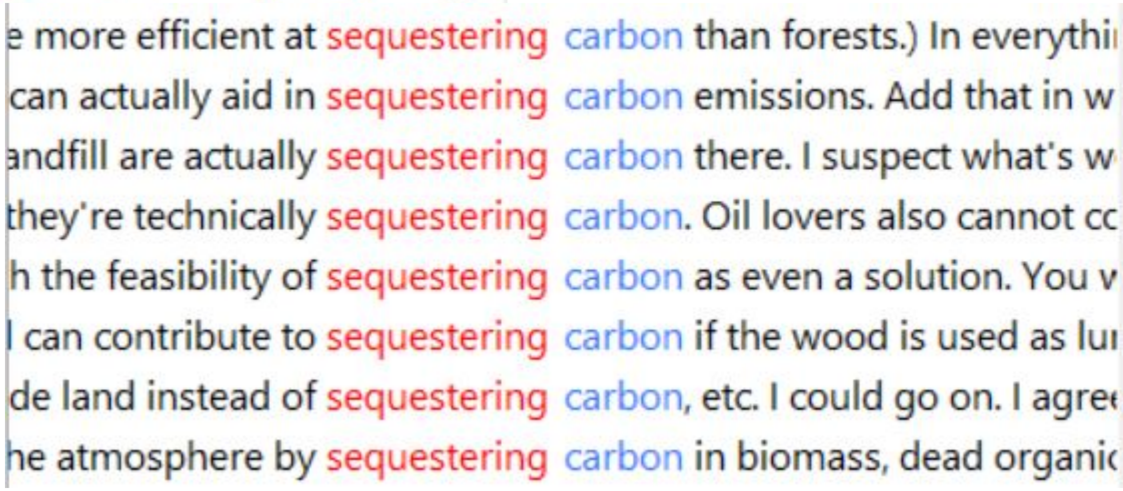
“I’m happy the US is leading the world in **carbon** emissions reductions, even after withdrawing from the globalist Paris Climate Accord. Go complain to communist China if you’re mad about **CO2** emissions.”

EXAMPLES OF CORPUS USAGE

2. Concordances analysis (AntConc, python)



```
 should be on the table. It's also carbon intensive. It's the only way to meet
too cost extensive and also very carbon intensive. The same guy complains a
 If there is no planned exit from carbon intensive assets there will be another
rivate citizens who demand the carbon intensive resources? They deserve it.
eaches co2 payoff. Metal is very carbon intensive to make. In San Francisco, a
iain has the potential to be very carbon intensive if progress isn't made quick
nd various ways of making less carbon intensive gas. Making the entire conc
be solved to lower emissions in carbon-intensive industries like steel and shi
```



```
e more efficient at sequestering carbon than forests.) In everythi
can actually aid in sequestering carbon emissions. Add that in w
andfill are actually sequestering carbon there. I suspect what's w
they're technically sequestering carbon. Oil lovers also cannot cc
h the feasibility of sequestering carbon as even a solution. You v
l can contribute to sequestering carbon if the wood is used as lu
de land instead of sequestering carbon, etc. I could go on. I agre
he atmosphere by sequestering carbon in biomass, dead organi
```

POS-tagger for Tweets

TweetNLP Java-based tokenizer and POS-tagger for tweets (http://www.cs.cmu.edu/~ark/TweetNLP/)

- Simple Python wrapper
- What for?

As part of our commitment to #FinancingAGreenFuture, we've worked with @thecarbontrust to better understand how customer spending activity during the lockdown affected the UK's carbon emissions across six key spending areas. Read the key findings here:... https://t.co/uk9Ipaszya

```
[('As', 'P', 0.898), ('part', 'N', 0.9962), ('of', 'P', 0.9984), ('our', 'D',
0.998), ('commitment', 'N', 0.9995), ('to', 'P', 0.9982),
('#FinancingAGreenFuture', '^', 0.8071), (',', ',', 0.9948), ("we've",
'L', 0.8994), ('worked', 'V', 0.9978), ('with', 'P', 0.9979),
('@thecarbontrust', '@', 0.9933), ('to', 'P', 0.9313), ('better', 'R',
0.4975), ('understand', 'V', 0.9431), ('how', 'R', 0.9793), ('customer',
'N', 0.548), ('spending', 'V', 0.964), ('activity', 'N', 0.9882), ('during',
'P', 0.9978), ('the', 'D', 0.9991), ('lockdown', 'N', 0.6142), ('affected',
'V', 0.9943), ('the', 'D', 0.9994), ("UK's", 'Z', 0.5561), ('carbon', 'N',
0.8667), ('emissions', 'N', 0.9973), ('across', 'P', 0.9988), ('six', '$',
0.5479), ('key', 'N', 0.6816), ('spending', 'V', 0.9804), ('areas', 'N',
0.9689), ('.', ',', 0.9978), ('Read', 'V', 0.99), ('the', 'D', 0.9996), ('key',
'N', 0.6015), ('findings', 'N', 0.9848), ('here', 'R', 0.9676), (':', ',',
0.9464), ('...', '~', 0.5559), ('https://t.co/uk9Ipaszya', 'U', 0.9965)]
```

NOT RELEVANT TOPICS

keywords : “..., carbon, carbon dioxide, ...”

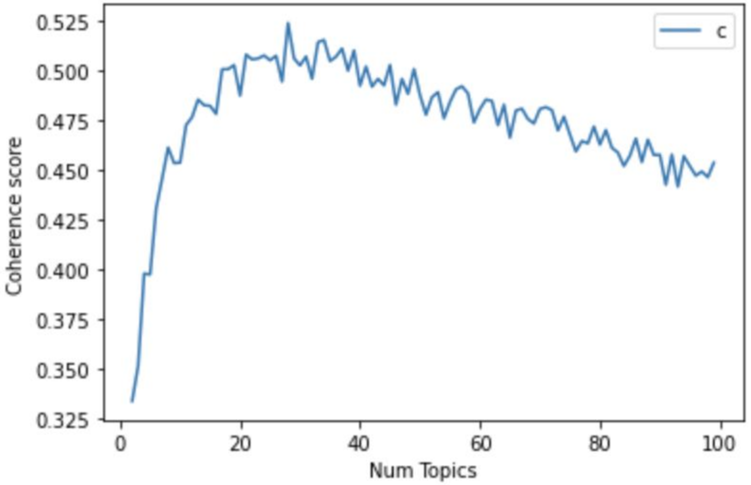
COVID context

> “If an old lady or young man have heart attack from reduced oxygen and increased **carbon dioxide** mask in your shop will you have doctor on standby to resuscitate.”

Topic Modeling with LDA

LDA: latent dirichlet allocation — automatic topic modeling

- Cleaning: stopwords, mentions, links, punctuation, smiles, etc.
- Input: Cleaned tweets, number of topics
  Output: dominant topic for each tweet
- Each topic is a combination of keywords with certain weightages



Coherence scores, 2-100 topics, Twitter Carbon subcorpus (1 827 40 tokens)

(0, '0.070*"carbon" + 0.039*"reduce" + 0.038*"emission" + 0.031*"build" + ' '0.025*"building"')

1 of 28 topics extracted from Twitter Carbon subcorpus, Coherence Value of 0.5239

TOPIC MODELING FOR CLEANING CORPUS

keywords : “..., carbon, carbon dioxide, ...”

COVID

[*dioxide, air, oxygen, mask, breathe, body, filter, people, face, wear*]

> “If an old lady or young man have heart attack from reduced oxygen and increased **carbon dioxide** mask in your shop will you have doctor on standby to resuscitate.”

Carbon monoxide poisoning

[*carbon, gas, monoxide, home, die, smoke, house, remember, family, alarm*]

> “the main take away from my 4 years of studying chemistry is that any of the fuel-burning appliances in your home can start producing **carbon** monoxide which can kill you overnight”

TOPIC MODELING FOR DETECTING SPECIFIC THEMES

American politics

[*people, stop, trump, science, biden, care, control, american, fly, lie*]

> “Well then you will be happy to know the **US** still **leads** the world in **carbon reduction** and **Trump** hasn’t started any new wars. For the first time in my life we aren’t trying to control other countries, instead **Trump** is working on peace deals and bringing troops home.”

Carbon as a fundamental chemical element

[*carbon, life, base, make, planet, end, human, live, time, save*]

> “**Carbon** neutrality is a lie. All **life depends on carbon dioxide**. If **life** was “**carbon neutral**”, **you’d no longer exist.** Plant life depends on CO2 for food. Stop the hoax - fossil fuels are not the evil it is reported to be.”

TWITTER CORPUS – CLEANING STEPS

TO DO:

1. Correct the spelling (thru, corbon dixoide, treesdon't, wtf) *Autospeller, TextBlob*
2. Filter non-English content - *TextBlob*

   "Sng je, if u trjaga xpa but still just bukak je tingkap. Carbon monoxide is produced from the partial oxidation of carbon-containing compounds..."

1. Remove semi-duplicates - *Locality Sensitive Hashing* (*LSH*)?

   "Hydrogen and Carbon Capture Tech Are Key To Net-Zero US Electricity, Study Says: The United States can generate affordable electricity without producing carbon dioxide emissions by 2035 by deploying hydrogen or carbon capture technology, accord..."

   "Hydrogen and Carbon Capture Tech Are Key To Net-Zero US Electricity, Study Says: The United States can generate affordable electricity without producing carbon dioxide emissions by 2035 by deploying hydrogen or carbon capture technology, according to a r..."