# Leveraging natural language processing to review health impacts of air pollution

**Nastassja A. Lewinski[1] and Bridget T. McInnes[2]**
1) Department of Chemical and Life Science Engineering, Virginia Commonwealth University, Richmond, VA
2) Department of Computer Science, Virginia Commonwealth University, Richmond, VA

## Research Questions

Here, we demonstrate the application of natural language processing (NLP) approaches to extracting information from journal articles describing experiments related to health effects of air pollution to determine how NLP can help with determining:
(1) What air pollutants are most studied?
(2) What health effects of air pollution are most prevalent?
(3) What biomarkers indicate exposure to air pollution?
(4) Where are gaps in air pollution research?
(4) How can NLP assist in database curation?

## Approach

We focused on four major components of a NLP pipeline, classification, topic identification, named entity recognition, relation detection.

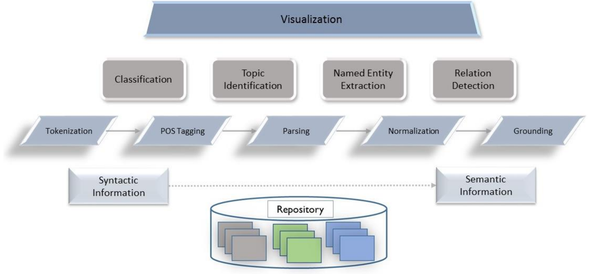


**Figure 1.** Generalized NLP workflow.

First, we started by performing a PubMed search using the KNIME Analytics Platform with the search term:
"air pollution" OR "air quality", AND ("human", "other animals", "animals").

The time domain of the search was January 1, 2002 to April 8, 2022 and the total number of documents retrieved after removing duplicates was 40,983.

The abstracts for each article were retrieved in plain text format for use in the NLP pipeline.

## Clustering

We used an existing multi-topic clustering tool, CLUTO (http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview), developed by the Karypis lab to divide the 40,983 papers into smaller groups based on cosine similarity. Two output sizes (40 and 100 clusters) were generated.

These document clusters primarily grouped the papers by the studied pollutant (e.g. radon, environmental tobacco smoke) followed by health outcome (e.g. asthma, cancer). Few groups (3 in the 40 cluster data set) represented papers correlating exposure to a specific pollutant with a specific health outcome. These three linked smoking with cancer, smoking with childhood asthma, and indoor air with sick building syndrome. When a larger number of smaller clusters was generated, groups of irrelevant papers (e.g. ear surgery, voice production) could be identified.

## Topic Identification

We used a custom tool, TopEx (http://topex.cctr.vcu.edu/), developed by the McInnes lab to identify subtopics within the clusters of documents.

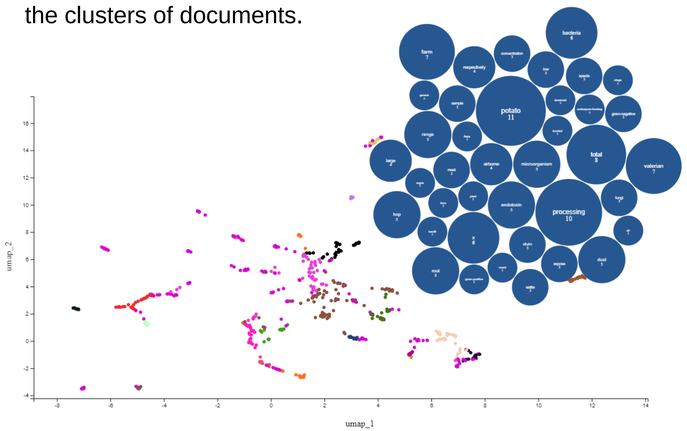


**Figure 2.** Representative **o**utput of TopEx clustering papers.

For example, the word cloud above is for a cluster of papers describing exposure to bacteria, bioaerosols, dust, allergens, and endotoxin. The words potato, processing, and farm suggest subtopic avenues for further exploration.

## Named Entity Recognition

We used a custom tool, Medacy (https://github.com/NLPatVCU/medaCy), developed by the McInnes lab to label four entities (species, cell line, chemical, and gene/protein). Because these entities are also included in the publicly available tool, PubTator (https://www.ncbi.nlm.nih.gov/research/pubtator/), we compared their performances.

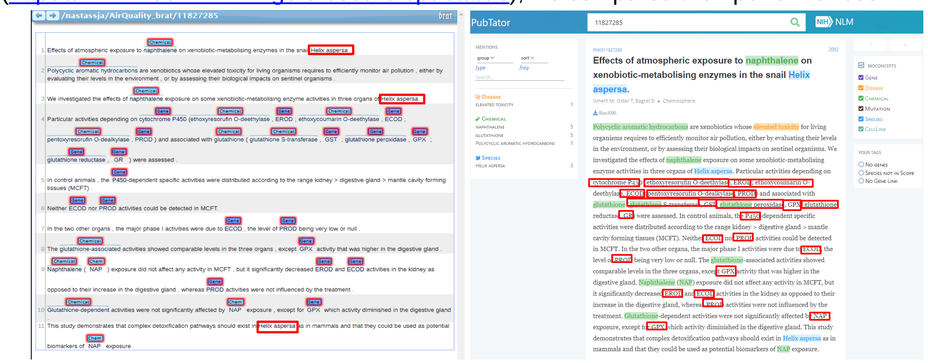


**Figure 3.** Representative output of Medacy (left) with PubTator (right) on the same abstract.

Overall Medacy identifies chemicals more completely while PubTator identifies species and cell line more completely. This is likely due to different data used to train each NER model. Proteins are labeled as both gene and chemical by Medacy while are not labeled at all by PubTator.

## Relation Detection

A custom relation detection system, RelEx (https://github.com/NLPatVCU/RelEx), has also been developed by the McInnes lab; however, the training data used to develop the model does not include relationships such as chemical-disease. It does include ADE-Drug; however, it is unlikely chemical-disease relationships can be identified using this model due to the different context.

## Reflections

From the perspective of performing a systematic review, some reflections on using these different NLP tools include:

- Clustering offers an additional approach to parring down data besides by keyword search. Well studied air pollutants and to a lesser extent prevalent related health effects can be identified from the clusters.
- Entity recognition identifies individual chemicals but not complex mixtures (e.g. particulate matter, tobacco smoke). Proteins present a challenge to entity recognition as they are labeled as both chemicals and genes which represents their roles as pollutants (e.g. endotoxin) and biomarkers (e.g. cytochrome P450).
- Cross comparing between documents is still not easy. Statistics both within a document and a document cluster are desired as well as structuring of entities into a database.

**Acknowledgements**
We would like to recognize students in the McInnes lab who developed TopEx (Dr. Amy Olex, Evan French), Medacy (Andriy Mulyar, Steele Farnsworth, Gabrielle Gurdin, Jorge Vargas, Neha Dill, Grant Matteo, Luke Maffey, Dr. Darshini Mahendran, Dr. Amy Olex), RelEx (Dr. Darshini Mahendran, Neha Dill), and Joshua Morriss for the KNIME search. This research was sponsored by Virginia Commonwealth University, the VCU Center for Clinical and Translational Research, and The Thomas F. and Kate Miller Jeffress Memorial Trust.

VCU College of Engineering