Chemical Entity Recognition System for English Literature User Guide

一、Introduction

1.1 Purpose

This system aims to quickly and conveniently identify chemical entities from environment-related English literature, establish a chemical list, and count their frequency of occurrence.

1.2 Significance

Chemical information reported widely in effective and reliable scientific literature can provide a basis for the assessment and screening of chemicals. However, the exponential growth and complexity of scientific literature make manual evaluation extremely challenging and time-consuming. This system retrieves information on over 900,000 chemicals from the highly relevant CompTox platform (https://comptox.epa.gov/dashboard/) and extends the chemical name list through the PubChem database (https://pubchem.ncbi.nlm.nih.gov/). By creating a multi-word dictionary of over 7.85 million chemical names, the system collects chemical names from English texts by matching word sequences formed after tokenization, establishing a chemical list, and counting their frequency to assist in the assessment and screening of chemicals in the environment.

1.3 Definitions

- System: The "Chemical Entity Recognition System for English Literature".
- User: Any individual who can use the system's functionalities.
- Literature: English literature published in English journals.
- Dictionary: The dictionary established by this system for matching chemical names in the text.

二、Overview

2.1 System Introduction

The system establishes three dictionaries containing a total of 7.85 million chemical names based on the number of words forming the chemical names (separated by spaces). It also supports additional dictionaries or the creation of new ones. By matching the word sequences formed after tokenizing English text with the dictionary, the system collects chemical names from the literature, establishes a chemical list, and counts the frequency of occurrence. The system realizes fast and convenient chemical entity recognition functions in English texts, which can be used for collecting chemical names from English literature.

2.2 System Environment

The system requires Windows 10 and Python 3.8 or above.

三、User Guide

3.1 System Introduction

The system mainly includes two parts: dictionary creation and named entity recognition. Dictionary creation includes two modules: creating a new dictionary and adding to an existing dictionary. Named entity recognition includes three modules: loading dictionaries, starting recognition, and outputting files.

Users can download all the files of this system from https://github.com/huangjiehui826/chemical_ner_v1. The main files include one Python file (chemical_NER.py or chemical_NER-English.py), nine text files (seven of which are stored in the initial folder, containing over 7.85 million chemical names for the next step of named entity recognition), a requirements.txt file for configuring the required Python libraries, and three pkl files stored in the initial folder, containing the basic information of over 7.85 million chemical names.

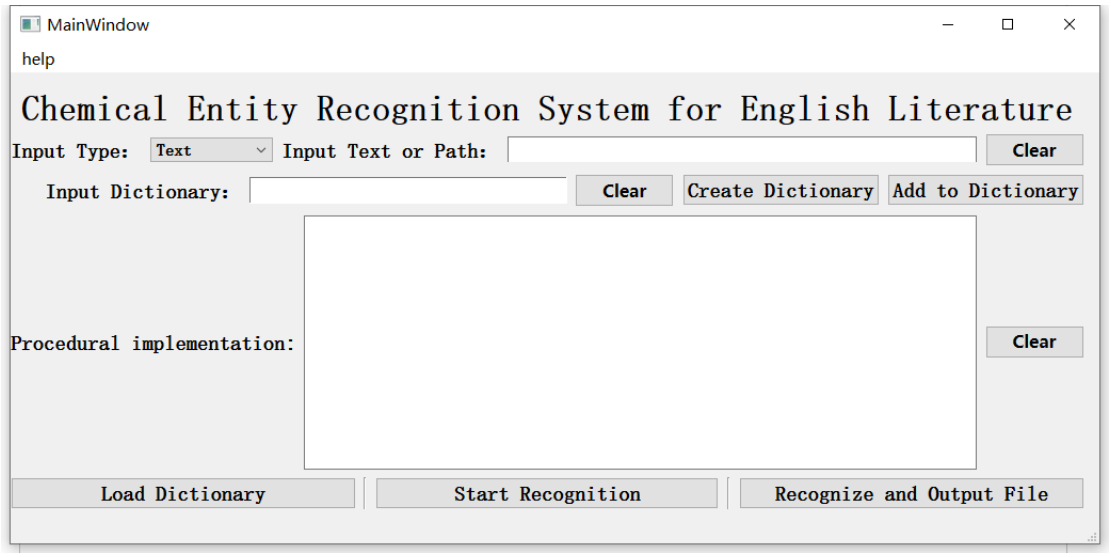


Figure 1 System Main Interface

3.2 Operating Steps

This section demonstrates the operating steps on Windows 10 using Anaconda3 (2021.05) and PyCharm (2021.2.2). Users need to install the corresponding software and set up the appropriate environment.

3.2.1 Load Environment and System

1. On Windows 10, open the command prompt by pressing "Win + R" and entering "cmd".
2. In the command prompt, create a new virtual environment named ner_demo by entering conda create -n ner_demo python=3.8.
3. Activate the newly created virtual environment with activate ner_demo and install the required Python libraries with
   pip install -r C:\Users\Desktop\chemical_NER_V1\requirements.txt.
4. Open chemical_NER.py in PyCharm, set the interpreter to ner_demo, and run the system.

3.2.2 Create a New Dictionary

1. Set "Input Type" to "Dictionary"

2. Enter the file path of the dictionary file containing the chemical names corresponding to CAS, SID, and CID in the "Input Text or Path" field
3. Enter the name of the dictionary series you want to set in the "Input Dictionary" field, default is "initial".
4. Click the "Create Dictionary" button.

3.2.3 Add to an Existing Dictionary

1. Set "Input Type" to "Dictionary".
2. Enter the file path of the additional dictionary file containing the chemical names corresponding to CAS, SID, and CID in the "Input Text or Path" field.
3. Enter the name of the dictionary series you want to add to in the "Input Dictionary" field, default is "initial"。
4. Click the "Add to Dictionary" button。

3.2.4 Load Dictionary

1. Set "Input Type" to "Dictionary"
2. Enter the name of the dictionary series you want to load in the "Input Dictionary" field, default is "initial"。
3. Click the "Load Dictionary" button。

3.2.5 Start Recognition

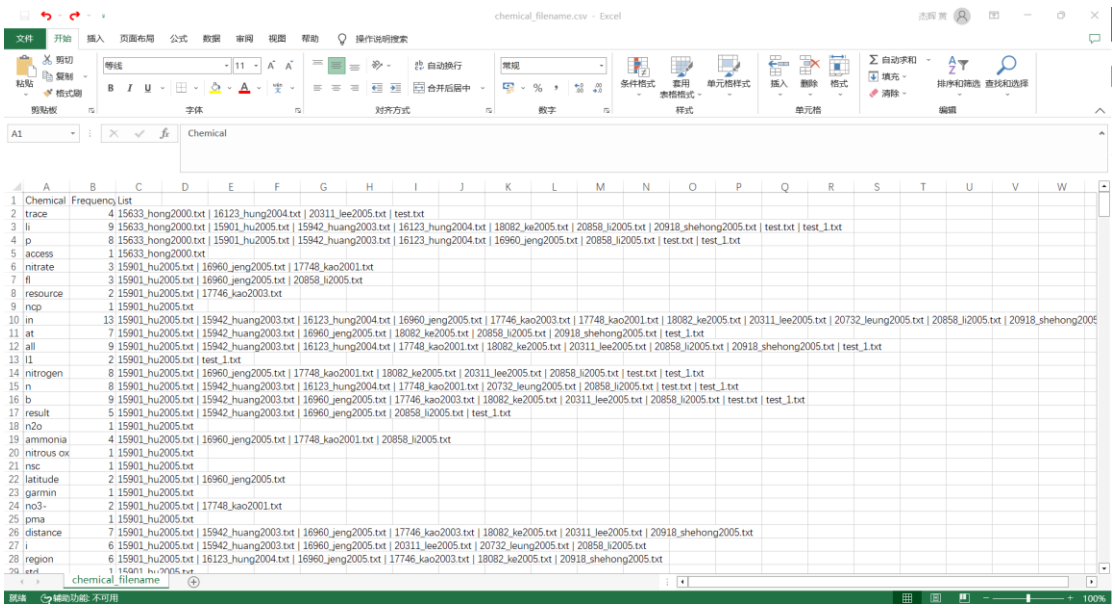
1. Set "Input Type" to "Text", "File", or "Folder".
2. Enter the text or file path you want to recognize in the "Input Text or Path" field.
3. Enter the name of the dictionary series you want to use in the "Input Dictionary" field, default is "initial".
4. Optionally, click the "Load Dictionary" button if you want to use a different dictionary series.
5. Click the "Start Recognition" button.

3.2.6 Recognize and Output File

1. Set "Input Type" to "Text", "File", or "Folder".
2. Enter the text or file path you want to recognize in the "Input Text or Path" field.
3. Enter the name of the dictionary series you want to use in the "Input Dictionary" field, default is "initial".
4. Optionally, click the "Load Dictionary" button if you want to use a different dictionary series.
5. Click the "Recognize and Output File" button.

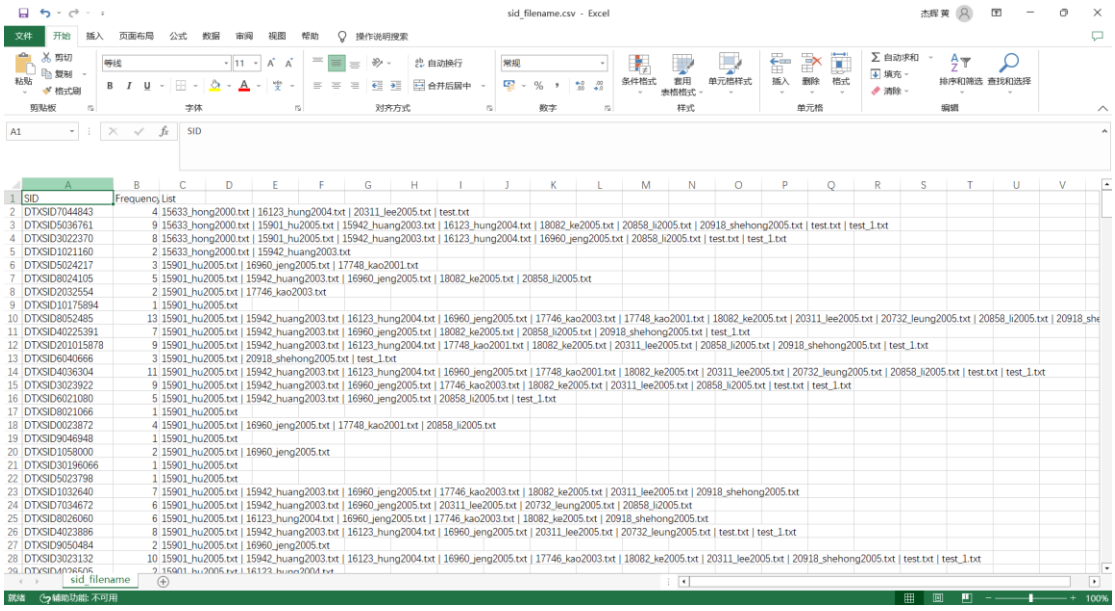
The system will automatically create an "output" folder in the system directory and generate a text file containing the recognized chemical names if text or file is entered, or generate three CSV files and three PKL files containing the chemical names and corresponding file names if a folder is entered.。

Figure 2 Reported Chemical Names（Input type "Text or File"）



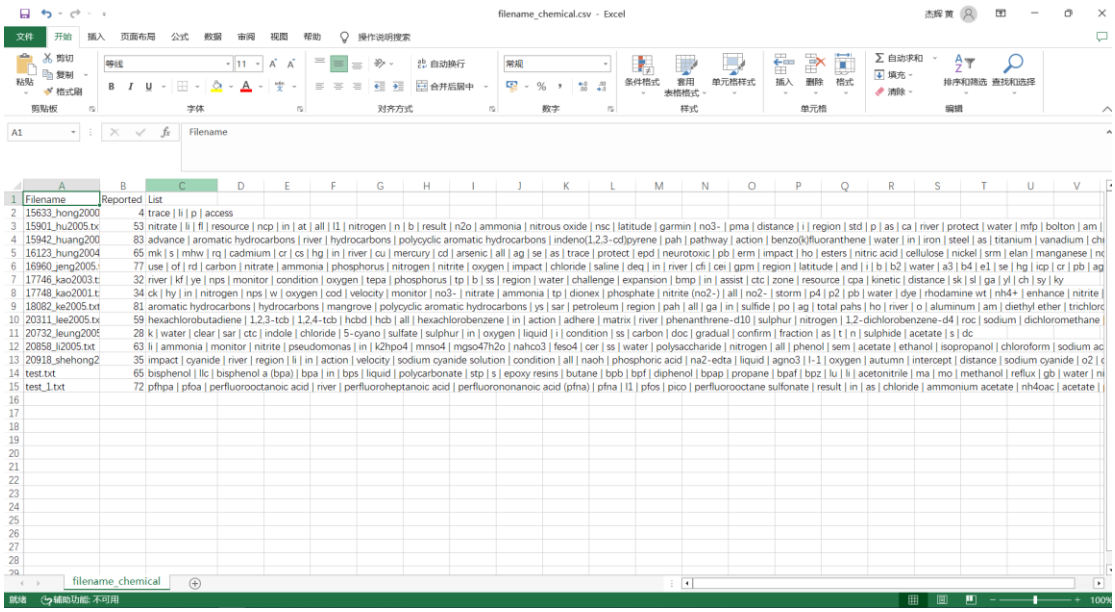
Chemical	Frequency	List
trace	4	15633_hong2000.txt | 16123_hung2004.txt | 20311_lee2005.txt | test.txt
li	9	15633_hong2000.txt | 15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 18082_ke2005.txt | 20858_li2005.txt | 20918_shehong2005.txt | test.txt | test_1.txt
p	8	15633_hong2000.txt | 15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 16960_jeng2005.txt | 20858_li2005.txt | test.txt | test_1.txt
access	1	15633_hong2000.txt
nitrate	3	15901_hu2005.txt | 16960_jeng2005.txt | 17748_kao2001.txt
fl	3	15901_hu2005.txt | 16960_jeng2005.txt | 20858_li2005.txt
resource	2	15901_hu2005.txt | 17746_kao2003.txt
ncp	1	15901_hu2005.txt
in	13	15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 16960_jeng2005.txt | 17746_kao2003.txt | 17748_kao2001.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20732_leung2005.txt | 20858_li2005.txt | 20918_shehong2005
at	7	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 18082_ke2005.txt | 20858_li2005.txt | 20918_shehong2005.txt | test_1.txt
all	9	15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 17748_kao2001.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20858_li2005.txt | 20918_shehong2005.txt | test_1.txt
l1	2	15901_hu2005.txt | test_1.txt
nitrogen	8	15901_hu2005.txt | 16960_jeng2005.txt | 17748_kao2001.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20858_li2005.txt | test.txt | test_1.txt
n	8	15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 17748_kao2001.txt | 20732_leung2005.txt | 20858_li2005.txt | test.txt | test_1.txt
b	9	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 17746_kao2003.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20858_li2005.txt | test.txt | test_1.txt
result	5	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 20858_li2005.txt | test_1.txt
n2o	1	15901_hu2005.txt
ammonia	4	15901_hu2005.txt | 16960_jeng2005.txt | 17748_kao2001.txt | 20858_li2005.txt
nitrous ox	1	15901_hu2005.txt
nsc	1	15901_hu2005.txt
latitude	2	15901_hu2005.txt | 16960_jeng2005.txt
garmin	1	15901_hu2005.txt
no3-	2	15901_hu2005.txt | 17748_kao2001.txt
pma	1	15901_hu2005.txt
distance	7	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 17746_kao2003.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20918_shehong2005.txt
i	6	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 20311_lee2005.txt | 20732_leung2005.txt | 20858_li2005.txt
region	6	15901_hu2005.txt | 16123_hung2004.txt | 16960_jeng2005.txt | 17746_kao2003.txt | 18082_ke2005.txt | 20918_shehong2005.txt

Figure 3 Names of chemicals reported and their corresponding file names



SID	Frequency	List
DTXSID7044843	4	15633_hong2000.txt | 16123_hung2004.txt | 20311_lee2005.txt | test.txt
DTXSID5036761	9	15633_hong2000.txt | 15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 18082_ke2005.txt | 20858_li2005.txt | 20918_shehong2005.txt | test.txt | test_1.txt
DTXSID3022370	8	15633_hong2000.txt | 15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 16960_jeng2005.txt | 20858_li2005.txt | test.txt | test_1.txt
DTXSID1021160	2	15633_hong2000.txt | 15942_huang2003.txt
DTXSID5024217	3	15901_hu2005.txt | 16960_jeng2005.txt | 17748_kao2001.txt
DTXSID8024105	5	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 18082_ke2005.txt | 20858_li2005.txt
DTXSID2032554	2	15901_hu2005.txt | 17746_kao2003.txt
DTXSID10175894	1	15901_hu2005.txt
DTXSID8052485	13	15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 16960_jeng2005.txt | 17746_kao2003.txt | 17748_kao2001.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20732_leung2005.txt | 20858_li2005.txt | 20918_she
DTXSID40225391	7	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 18082_ke2005.txt | 20858_li2005.txt | 20918_shehong2005.txt | test_1.txt
DTXSID201015878	9	15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 17748_kao2001.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20858_li2005.txt | 20918_shehong2005.txt | test_1.txt
DTXSID6040666	3	15901_hu2005.txt | 20918_shehong2005.txt | test_1.txt
DTXSID4036304	11	15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 16960_jeng2005.txt | 17748_kao2001.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20732_leung2005.txt | 20858_li2005.txt | test.txt | test_1.txt
DTXSID3023922	9	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 17746_kao2003.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20858_li2005.txt | test.txt | test_1.txt
DTXSID6021080	5	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 20858_li2005.txt | test_1.txt
DTXSID8021066	1	15901_hu2005.txt
DTXSID0023872	4	15901_hu2005.txt | 16960_jeng2005.txt | 17748_kao2001.txt | 20858_li2005.txt
DTXSID9046948	1	15901_hu2005.txt
DTXSID1058000	2	15901_hu2005.txt | 16960_jeng2005.txt
DTXSID30196066	1	15901_hu2005.txt
DTXSID5023798	1	15901_hu2005.txt
DTXSID1032640	7	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 17746_kao2003.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20918_shehong2005.txt
DTXSID7034672	6	15901_hu2005.txt | 15942_huang2003.txt | 16960_jeng2005.txt | 20311_lee2005.txt | 20732_leung2005.txt | 20858_li2005.txt
DTXSID8026060	6	15901_hu2005.txt | 16123_hung2004.txt | 16960_jeng2005.txt | 17746_kao2003.txt | 18082_ke2005.txt | 20918_shehong2005.txt
DTXSID4023886	8	15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 16960_jeng2005.txt | 20311_lee2005.txt | 20732_leung2005.txt | test.txt | test_1.txt
DTXSID9050484	2	15901_hu2005.txt | 16960_jeng2005.txt
DTXSID3023132	10	15901_hu2005.txt | 15942_huang2003.txt | 16123_hung2004.txt | 16960_jeng2005.txt | 17746_kao2003.txt | 18082_ke2005.txt | 20311_lee2005.txt | 20918_shehong2005.txt | test.txt | test_1.txt

Figure 4 Reported sid and its corresponding frequency and file name



Filename	Reported	List
15633_hong2000	4	trace | li | p | access
15901_hu2005.tx	53	nitrate | li | fl | resource | ncp | in | at | all | l1 | nitrogen | n | b | result | n2o | ammonia | nitrous oxide | nsc | latitude | garmin | no3- | pma | distance | i | region | std | p | as | ca | river | protect | water | mfp | bolton | am |
15942_huang200	83	advance | aromatic hydrocarbons | river | hydrocarbons | polycyclic aromatic hydrocarbons | indeno(1,2,3-cd)pyrene | pah | pathway | action | benzo(k)fluoranthene | water | in | iron | steel | as | titanium | vanadium | ch
16123_hung2004	65	mk | s | mhw | rq | cadmium | cr | cs | hg | in | river | cu | mercury | cd | arsenic | all | ag | se | as | trace | protect | epd | neurotoxic | pb | erm | impact | ho | esters | nitric acid | cellulose | nickel | srm | elan | manganese | n
16960_jeng2005.	77	use | of | rd | carbon | nitrate | ammonia | phosphorus | nitrogen | nitrite | oxygen | impact | chloride | saline | deq | in | river | cfi | cei | gpm | region | latitude | and | i | b | b2 | water | a3 | b4 | e1 | se | hg | icp | cr | pb | ag
17746_kao2003.t	32	river | kf | ye | nps | monitor | condition | oxygen | tepa | phosphorus | tp | b | ss | region | water | challenge | expansion | bmp | in | assist | ctc | zone | resource | cpa | kinetic | distance | sk | sl | ga | yl | ch | sy | ky
17748_kao2001.t	34	ck | hy | in | nitrogen | nps | w | oxygen | cod | velocity | monitor | no3- | nitrate | ammonia | tp | dionex | phosphate | nitrite (no2-) | all | no2- | storm | p4 | p2 | pb | water | dye | rhodamine wt | nh4+ | enhance | nitrite |
18082_ke2005.txt	81	aromatic hydrocarbons | hydrocarbons | mangrove | polycyclic aromatic hydrocarbons | ys | sar | petroleum | region | pah | all | ga | in | sulfide | po | ag | total pahs | ho | river | o | aluminum | am | diethyl ether | trichlorc
20311_lee2005.tx	59	hexachlorobutadiene | 1,2,3-tcb | 1,2,4-tcb | hcbd | hcb | all | hexachlorobenzene | in | action | adhere | matrix | river | phenanthrene-d10 | sulphur | nitrogen | 1,2-dichlorobenzene-d4 | roc | sodium | dichloromethane |
20732_leung200	28	k | water | clear | sar | ctc | indole | chloride | 5-cyano | sulfate | sulphur | in | oxygen | liquid | i | condition | ss | carbon | doc | gradual | confirm | fraction | as | t | n | sulphide | acetate | s | dc
20858_li2005.txt	63	li | ammonia | monitor | nitrite | pseudomonas | in | k2hpo4 | mnso4 | mgso47h2o | nahco3 | feso4 | cer | ss | water | polysaccharide | nitrogen | all | phenol | sem | acetate | ethanol | isopropanol | chloroform | sodium ac
20918_shehong2	35	impact | cyanide | river | region | li | in | action | velocity | sodium cyanide solution | condition | all | naoh | phosphoric acid | na2-edta | liquid | agno3 | l-1 | oxygen | autumn | intercept | distance | sodium cyanide | o2 |
test.txt	65	bisphenol | llc | bisphenol a (bpa) | bpa | in | bps | liquid | polycarbonate | stp | s | epoxy resins | butane | bpb | bpf | diphenol | bpap | propane | bpaf | bpz | lu | li | acetonitrile | ma | mo | methanol | reflux | gb | water | ni
test_1.txt	72	pfhpa | pfoa | perfluorooctanoic acid | river | perfluoroheptanoic acid | perfluorononanoic acid (pfna) | pfna | l1 | pfos | pico | perfluorooctane sulfonate | result | in | as | chloride | ammonium acetate | nh4oac | acetate |

Figure 5 Names of chemicals reported in the document