In [1]:
```
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Load Dataset

In [2]:
```
file_names = os.listdir("Dataset/SemEval2017/docsutf8/")

texts = []
keys = []

file_path = "Dataset/SemEval2017/docsutf8/"
key_path = "Dataset/SemEval2017/keys/"

for file_name in file_names:
    key_name = file_name.split(".")[0] + ".key"

    with open(file_path + file_name, encoding="utf8") as f:
        text = f.read()
        texts.append(text)

    with open(key_path + key_name, encoding="utf8") as f:
        key = f.readlines()
        key = [txt[:-1] for txt in key]
        key = ", ".join(key)
        keys.append(key)


df = pd.DataFrame({"Text" : texts, "Keys" : keys})
df.head()
```

Out[2]:

	Text	Keys
0	Complex Langevin (CL) dynamics [1,2] provides...	CL, complexified configuration space, Complex ...
1	Nuclear theory devoted major efforts since 4 d...	C60, combining quantum features, field of clus...
2	The next important step might be the derivatio...	continuum space-time, Dirac equation, future r...
3	This work shows how our approach based on the ...	class virial expansions, field partition funct...
4	A fluctuating vacuum is a general feature of q...	a collection of fermionic fields describing co...

Rake [Rapid Automatic Keyword Extraction algorithm]

In [3]:
```
from rake_nltk import Rake

def rake_top_k(text, k=10):

    r = Rake()
    r.extract_keywords_from_text(text)

    result = r.get_ranked_phrases()[:k]
    return result
```

Using Rake

In [4]:
```
rake_top_k(df["Text"][100], 20)
```

Out[4]:
```
['activation energies calculated using empirical pair potentials',
 'activation energies calculated using dft',
 'examined three different potentials',
 'fuel matrix initially accommodated',
 'jahn – teller distortion',
 'point defects trap sites',
 'facilitate net xe diffusion',
 '6 – 8 ].',
 'defect trap sites',
 'activation energies',
 '‘ hop ’',
 'schottky trivacancy defects',
 'rate determining step',
 'govers et al',
 'different stoichiometric regimes',
 '11 ]) coupled',
 '– xe interactions',
 'u – xe',
 '15 – 7',
 'vary strongly depending']
```

Ground Truth

In [5]:
```
df["Keys"][100]
```

Out[5]:
```
'Activation, bubbles, charge, crystal, crystal stoichiometry, defect trap sites, DFT, diffusion, Diffusion, empirical pair potentials, fission, fuel matrix, gas atom, grain boundaries, ‘hop’ into, Jahn–Teller distortion, loop around, migration, noble gas atoms, O–Xe, point defects trap sites, potential, potentials, rearrangement, Schottky trivacancy defects, UO2, UO2+x, UO2−x, uranium, U–Xe, VU, VU defect, Xe, Xe diffusion'
```