Titel der Arbeit: Development of a solution for genetic analysis of ALL genomes by implementing Late Allocation

Name: Jacqueline Franßen

Matrikelnummer: 496804

Data Preparation ¶

- Imports (Pandas, Numpy, Spacy, NLTK, Gensim)
- · Create a dataframe object from given dataset
- · Filter Dataframe for relevant data for LDA

In [1]:

```
import pandas as pd
import numpy as np
import re
import spacy
import nltk
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
nltk.download('wordnet')
from nltk.corpus import stopwords
nltk.download('stopwords')
from nltk.tokenize import word_tokenize
nltk.download('punkt')
from nltk import word_tokenize,sent_tokenize
from nltk.corpus import wordnet as wn
from nltk.stem.wordnet import WordNetLemmatizer
import random
from gensim import corpora
import pickle
import gensim
from gensim.models import CoherenceModel
from gensim.test.utils import common texts
from gensim.corpora.dictionary import Dictionary
import pyLDAvis.gensim
[nltk_data] Downloading package wordnet to
[nltk_data]
                C:\Users\jacky\AppData\Roaming\nltk_data...
[nltk_data]
              Package wordnet is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]
               C:\Users\jacky\AppData\Roaming\nltk_data...
[nltk_data]
              Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]
                C:\Users\jacky\AppData\Roaming\nltk_data...
[nltk_data]
              Package punkt is already up-to-date!
```

Data import

Create a dataframe object from given dataset

```
In [2]:
```

```
df1 = pd.read_csv('all_genes.txt', error_bad_lines=False, sep='\t', comment='#', low_memory=False, header=0)
df1.head()
df1
```

Out[2]:

	tax_id	Org_name	GenelD	CurrentID	Status	Symbol	Aliases	description	other_designations	map_location	chromosome ge
0	9606	Homo sapiens	10320	0	live	IKZF1	CVID13, Hs.54452, IK1, IKAROS, LYF1, LyF-1, PP	IKAROS family zinc finger 1	DNA-binding protein Ikaros CLL-associated anti	7p12.2	7
1	9606	Homo sapiens	84159	0	live	ARID5B	DESRT, MRF-2, MRF2	AT-rich interaction domain 5B	AT-rich interactive domain-containing protein	10q21.2	10
2	9606	Homo sapiens	1053	0	live	CEBPE	C/EBP-epsilon, CRP1	CCAAT enhancer binding protein epsilon	CCAAT/enhancer- binding protein epsilon CCAAT/e	14q11.2	14
3	9606	Homo sapiens	1644	0	live	DDC	AADC	dopa decarboxylase	aromatic-L-amino-acid decarboxylase dopa decar	7p12.2-p12.1	7
4	9606	Homo sapiens	7102	0	live	TSPAN7	A15, CCG-B7, CD231, DXS1692E, MRX58, MXS1, TAL	tetraspanin 7	tetraspanin-7 CD231 antigen T-cell acute lymph	Xp11.4	х
5	9606	Homo sapiens	8009	0	live	LALL	NaN	Lymphomatous acute lymphoblastic leukemia	NaN	9p22-p21	9
6	9606	Homo sapiens	101928483	0	live	NALT1	LINC01573, MIR4674HG, NALT, TCONS_I2_00029132	NOTCH1 associated IncRNA in T cell acute lymph	MIR4674 host gene (non-protein coding) MIR4674	9q34.3	9
7	9606	Homo sapiens	338436	0	live	BLACE	NaN	B cell acute lymphoblastic leukemia expressed	NaN	7q36.3	7
8	9606	Homo sapiens	1029	0	live	CDKN2A	ARF, CDK4I, CDKN2, CMM2, INK4, INK4A, MLM, MTS	cyclin dependent kinase inhibitor 2A	cyclin-dependent kinase inhibitor 2A CDK4 inhi	9p21.3	9
9	9606	Homo sapiens	5142	0	live	PDE4B	DPDE4, PDEIVB	phosphodiesterase 4B	cAMP-specific 3',5'- cyclic phosphodiesterase 4	1p31.3	1
10	9606	Homo sapiens	3600	0	live	IL15	IL-15	interleukin 15	interleukin-15	4q31.21	4
11	9606	Homo sapiens	8626	0	live	TP63	AIS, B(p51A), B(p51B), EEC3, KET, LMS, NBP, OF	tumor protein p63	tumor protein 63 amplified in squamous cell ca	3q28	3
12	9606	Homo sapiens	2625	0	live	GATA3	HDR, HDRS	GATA binding protein 3	trans-acting T-cell- specific transcription fac	10p14	10
13	9606	Homo sapiens	8202	0	live	NCOA3	ACTR, AIB-1, AIB1, CAGH16, CTG26, KAT13B, RAC3	nuclear receptor coactivator 3	nuclear receptor coactivator 3 CBP- interacting	20q13.12	20
14	9606	Homo sapiens	6262	0	live	RYR2	ARVC2, ARVD2, RYR-2, RyR, VTSIP	ryanodine receptor 2	ryanodine receptor 2 cardiac muscle ryanodine	1q43	1
15	9606	Homo sapiens	3738	0	live	KCNA3	HGK5, HLK3, HPCN3, HUKIII, KV1.3, MK3, PCN3	potassium voltage- gated channel subfamily A me	potassium voltage- gated channel subfamily A me	1p13.3	1
16	9606	Homo sapiens	5305	0	live	PIP4K2A	PI5P4KA, PIP5K2A, PIP5KII-alpha, PIP5KIIA, PIPK	phosphatidylinositol- 5-phosphate 4- kinase type	phosphatidylinositol 5- phosphate 4-kinase type	10p12.2	10
17	9606	Homo sapiens	84441	0	live	MAML2	MAM-3, MAM2, MAM3, MLL-MAML2	mastermind like transcriptional coactivator 2	mastermind-like protein 2 mam- 2 mastermind-like 2	11q21	11
18	9606	Homo sapiens	5795	0	live	PTPRJ	CD148, DEP1, HPTPeta, R-PTP- ETA, SCC1	protein tyrosine phosphatase receptor type J	receptor-type tyrosine- protein phosphatase eta	11p11.2	11
19	9606	Homo sapiens	56288	0	live	PARD3	ASIP, Baz, PAR3, PAR3alpha, PARD- 3A, PPP1R118,	par-3 family cell polarity regulator	partitioning defective 3 homolog CTCL tumor an	10p11.22- p11.21	10
20	9606	Homo sapiens	59350	0	live	RXFP1	LGR7, RXFPR1	relaxin family peptide receptor 1	relaxin receptor 1 leucine-rich repeat- contain	4q32.1	4
21	9606	Homo sapiens	6916	0	live	TBXAS1	BDPLT14, CYP5, CYP5A1, GHOSAL, THAS, TS, TXAS,	thromboxane A synthase 1	thromboxane-A synthase TXA synthase cytochrome	7q34	7
22	9606	Homo sapiens	2887	0	live	GRB10	GRB-IR, Grb-10, IRBP, MEG1, RSS	growth factor receptor bound protein 10	growth factor receptor- bound protein 10 GRB10	7p12.1	7
23	9606	Homo sapiens	9844	0	live	ELMO1	CED-12, CED12, ELMO-1	engulfment and cell motility 1	engulfment and cell motility protein 1 ced- 12	7p14.2-p14.1	7
24	9606	Homo sapiens	7157	0	live	TP53	BCC7, BMFS5, LFS1, P53, TRP53	tumor protein p53	cellular tumor antigen p53 antigen NY-CO- 13 mu	17p13.1	17
25	9606	Homo sapiens	9863	0	live	MAGI2	ACVRIP1, AIP-1, AIP1, ARIP1, MAGI-2, NPHS15, S	membrane associated guanylate kinase, WW and P	membrane-associated guanylate kinase, WW and P	7q21.11	7
26	9606	Homo sapiens	2952	0	live	GSTT1	NaN	glutathione S- transferase theta 1	glutathione S- transferase theta- 1 GST class-th	22q11.23	22

	tax_id	Org_name	GenelD	CurrentID	Status	Symbol	Aliases	description	other_designations	map_location	chromosome g
27	9606	Homo sapiens	6014	0	live	RIT2	RIBA, RIN, ROC2	Ras like without CAAX 2	GTP-binding protein Rit2 GTP-binding protein R	18q12.3	18
28	9606	Homo sapiens	340419	0	live	RSPO2	CRISTIN2, HHRRD, TETAMS2	R-spondin 2	R-spondin-2 R-spondin 2 homolog roof plate- spe	8q23.1	8
29	9606	Homo sapiens	5836	0	live	PYGL	GSD6	glycogen phosphorylase L	glycogen phosphorylase, liver form phosphoryla	14q22.1	14
562	9606	Homo sapiens	127933	0	live	UHMK1	KIS, KIST, P-CIP2	U2AF homology motif kinase 1	serine/threonine- protein kinase Kist KIS prote	1q23.3	1
563	9606	Homo sapiens	388585	0	live	HES5	bHLHb38	hes family bHLH transcription factor 5	transcription factor HES-5 class B basic helix	1p36.32	1
564	9606	Homo sapiens	55729	0	live	ATF7IP	AM, ATF-IP, MCAF, MCAF1, p621	activating transcription factor 7 interacting	activating transcription factor 7-interacting	12p13.1	12
565	9606	Homo sapiens	574504	0	live	MIR502	MIRN502, hsa-mir- 502, mir-502	microRNA 502	NaN	Xp11.23	Х
566	9606	Homo sapiens	554212	0	live	MIR448	MIRN448, hsa-mir- 448, miRNA448	microRNA 448	NaN	Xq23	Х
567	9606	Homo sapiens	83937	0	live	RASSF4	AD037	Ras association domain family member 4	ras association domain-containing protein 4 Ra	10q11.21	10
568	9606	Homo sapiens	414899	0	live	BLID	BRCC2	BH3-like motif containing, cell death inducer	BH3-like motif- containing cell death inducer b	11q24.1	11
569	9606	Homo sapiens	79370	0	live	BCL2L14	BCLG	BCL2 like 14	apoptosis facilitator Bcl-2-like protein 14 BC	12p13.2	12
570	9606	Homo sapiens	79368	0	live	FCRL2	CD307b, FCRH2, IFGP4, IRTA4, SPAP1, SPAP1A, SP	Fc receptor like 2	Fc receptor-like protein 2 IFGP family protein	1q23.1	1
571	9606	Homo sapiens	266977	0	live	ADGRF1	GPR110, KPG_012, PGR19, hGPCR36	adhesion G protein- coupled receptor F1	adhesion G-protein coupled receptor F1 G prote	6	6
572	9606	Homo sapiens	92912	0	live	UBE2Q2	NaN	ubiquitin conjugating enzyme E2 Q2	ubiquitin-conjugating enzyme E2 Q2 E2 ubiquiti	15q24.2	15
573	9606	Homo sapiens	115350	0	live	FCRL1	CD307a, FCRH1, IFGP1, IRTA5	Fc receptor like 1	Fc receptor-like protein 1 IFGP family protein	1q23.1	1
574	9606	Homo sapiens	80206	0	live	FHOD3	FHOS2, Formactin2	formin homology 2 domain containing 3	FH1/FH2 domain- containing protein 3 formactin	18q12.2	18
575	9606	Homo sapiens	54970	0	live	TTC12	TPARM	tetratricopeptide repeat domain 12	tetratricopeptide repeat protein 12 TPR repeat	11q23.2	11
576	9606	Homo sapiens	6887	0	live	TAL2	NaN	TAL bHLH transcription factor 2	T-cell acute lymphocytic leukemia protein 2 T	9q31.2	9
577	9606	Homo sapiens	693197	0	live	MIR612	MIRN612, hsa-mir- 612	microRNA 612	NaN	11q13.1	11
578	9606	Homo sapiens	1389	0	live	CREBL2	NaN	cAMP responsive element binding protein like 2	cAMP-responsive element-binding protein-like 2	12p13.1	12
579	9606	Homo sapiens	160365	0	live	CLECL1	DCAL-1, DCAL1	C-type lectin like 1	C-type lectin-like domain family 1 DC- associat	12p13.31	12
580	9606	Homo sapiens	128710	0	live	SLX4IP	C20orf94, bA204H22.1, bA254M13.1, dJ1099D15.3	SLX4 interacting protein	protein SLX4IP	20p12.2	20
581	9606	Homo sapiens	57824	0	live	HMHB1	HB-1, HB-1Y, HLA- HB1	histocompatibility minor HB-1	minor histocompatibility protein HB-1 minor hi	5q31.3	5
582	9606	Homo sapiens	162979	0	live	ZNF296	ZFP296, ZNF342	zinc finger protein 296	zinc finger protein 296 zinc finger protein 342	19q13.32	19
583	9606	Homo sapiens	574493	0	live	MIR520H	MIRN520H	microRNA 520h	hsa-mir-520h	19q13.42	19
584	9606	Homo sapiens	92241	0	live	RCSD1	CAPZIP, MK2S4	RCSD domain containing 1	capZ-interacting protein RCSD domain- containin	1q24.2	1
585	9606	Homo sapiens	107980440	0	live	LOC107980440	NaN	ABL breakpoint recombination region	ABL proto-oncogene 1, non-receptor tyrosine ki	9q34.1	9
586	9606	Homo sapiens	107963955	0	live	LOC107963955	NaN	BCR-ABL major- breakpoint cluster region	BCR p210 Philadelphia chromosome recombination	22q11.23	22

	tax_id	Org_name	GeneID	CurrentID	Status	Symbol	Aliases	description	other_designations	map_location	chromosome	ge
587	9606	Homo sapiens	107963951	0	live	LOC107963951	NaN	BCR-ABL minor- breakpoint cluster region	BCR p190 Philadelphia chromosome recombination	22q11.23	22	
588	9606	Homo sapiens	192343	0	live	NEWENTRY	NaN	Record to support submission of GeneRIFs for a	NaN	NaN	NaN	
589	9606	Homo sapiens	107648866	0	live	LOC107648866	NaN	meiotic recombination hotspot DNA3	NaN	6p21.3	6	
590	9606	Homo sapiens	3197	0	live	HOXA@	HOX1@	homeobox A cluster	homeo box A cluster	7p15.2	7	
591	129875	Human mastadenovirus A	1460850	0	live	E2A	HAdVAgp12	single-stranded DNA-binding protein	single-stranded DNA- binding protein	NaN	NaN	

592 rows × 18 columns

Only get relevant information from dataframe (2 colummns: Symbol and other_designations)

Filter Dataframe for relevant data for LDA

```
In [3]:
```

```
header = 0)
df2 = df2.dropna() # remove NaN values
# concat the 2 columns to one
df2 = df2['Symbol'].astype(str) + ' ' + df2['other_designations']
df2
Out[3]:
0
       IKZF1 DNA-binding protein Ikaros|CLL-associate...
       ARID5B AT-rich interactive domain-containing p...
1
       CEBPE CCAAT/enhancer-binding protein epsilon C...
2
       DDC aromatic-L-amino-acid decarboxylase dopa d...
       TSPAN7 tetraspanin-7 CD231 antigen T-cell acut...
6
       NALT1 MIR4674 host gene (non-protein coding) | M...
       CDKN2A cyclin-dependent kinase inhibitor 2A CD...
       PDE4B cAMP-specific 3',5'-cyclic phosphodieste...
                                     IL15 interleukin-15
10
       TP63 tumor protein 63 amplified in squamous ce...
11
12
       GATA3 trans-acting T-cell-specific transcripti...
       NCOA3 nuclear receptor coactivator 3 CBP-inter...
13
       RYR2 ryanodine receptor 2 cardiac muscle ryano...
14
15
       KCNA3 potassium voltage-gated channel subfamil...
       PIP4K2A phosphatidylinositol 5-phosphate 4-kin...
16
       MAML2 mastermind-like protein 2|mam-2|mastermi...
17
18
       PTPRJ receptor-type tyrosine-protein phosphata...
       PARD3 partitioning defective 3 homolog|CTCL tu...
19
20
       RXFP1 relaxin receptor 1 leucine-rich repeat-c...
       TBXAS1 thromboxane-A synthase|TXA synthase|cyt...
21
22
       GRB10 growth factor receptor-bound protein 10 | ...
23
       ELMO1 engulfment and cell motility protein 1 \mid c...
24
       TP53 cellular tumor antigen p53|antigen NY-CO-...
25
       MAGI2 membrane-associated guanylate kinase, WW...
26
       GSTT1 glutathione S-transferase theta-1|GST cl...
27
       RIT2 GTP-binding protein Rit2|GTP-binding prot...
28
       RSPO2 R-spondin-2|R-spondin 2 homolog|roof pla...
29
       PYGL glycogen phosphorylase, liver form phosph...
30
       MTHFR methylenetetrahydrofolate reductase | 5,10...
       KCNE4 potassium voltage-gated channel subfamil...
31
       SIAE sialate O-acetylesterase|H-Lse|cytosolic ...
557
558
       MLXIP MLX-interacting protein | Mlx interactor | c...
       TSGA10 testis-specific gene 10 protein cancer/...
559
560
       DPH1 2-(3-amino-3-carboxypropyl)histidine synt...
       TPD52L2 tumor protein D54 HCCR-binding protein...
561
562
       UHMK1 serine/threonine-protein kinase Kist|KIS...
563
       HES5 transcription factor HES-5 class B basic ...
564
       ATF7IP activating transcription factor 7-inter...
567
       RASSF4 ras association domain-containing prote...
       BLID BH3-like motif-containing cell death indu...
568
       BCL2L14 apoptosis facilitator Bcl-2-like prote...
569
570
       FCRL2 Fc receptor-like protein 2 IFGP family p...
       ADGRF1 adhesion G-protein coupled receptor F1|...
571
       UBE2Q2 ubiquitin-conjugating enzyme E2 Q2 E2 u...
572
       FCRL1 Fc receptor-like protein 1|IFGP family p...
573
574
       FHOD3 FH1/FH2 domain-containing protein 3|form...
       TTC12 tetratricopeptide repeat protein 12|TPR ...
575
       TAL2 T-cell acute lymphocytic leukemia protein...
576
       CREBL2 cAMP-responsive element-binding protein...
578
579
       CLECL1 C-type lectin-like domain family 1|DC-a...
                                   SLX4IP protein SLX4IP
580
       HMHB1 minor histocompatibility protein HB-1|mi...
581
582
       ZNF296 zinc finger protein 296 zinc finger pro...
583
                                    MIR520H hsa-mir-520h
584
       RCSD1 capZ-interacting protein RCSD domain-con...
585
       LOC107980440 ABL proto-oncogene 1, non-recepto...
586
       LOC107963955 BCR p210 Philadelphia chromosome ...
587
       LOC107963951 BCR p190 Philadelphia chromosome ...
590
                               HOXA@ homeo box A cluster
                 E2A single-stranded DNA-binding protein
Length: 573, dtype: object
```

Data Preparation

- Tokenization
- · remove special characters and numbers
- · remove stopwords
- · define methods for bigrams and trigrams

In [4]:

Tokenize

```
df2 tokens = df2.apply(word tokenize)
print(df2_tokens)
        [IKZF1, DNA-binding, protein, Ikaros|CLL-assoc...
        [ARID5B, AT-rich, interactive, domain-containi...
1
2
        [CEBPE, CCAAT/enhancer-binding, protein, epsil...
3
        [DDC, aromatic-L-amino-acid, decarboxylase|dop...
4
        [TSPAN7, tetraspanin-7|CD231, antigen|T-cell, ...
        [NALT1, MIR4674, host, gene, (, non-protein, c...
[CDKN2A, cyclin-dependent, kinase, inhibitor, ...
6
8
9
        [PDE4B, cAMP-specific, 3',5'-cyclic, phosphodi...
                                     [IL15, interleukin-15]
        [TP63, tumor, protein, 63 amplified, in, squam...
11
        [GATA3, trans-acting, T-cell-specific, transcr...
12
13
        [NCOA3, nuclear, receptor, coactivator, 3|CBP-...
        [RYR2, ryanodine, receptor, 2|cardiac, muscle,...
14
        [KCNA3, potassium, voltage-gated, channel, sub...
15
        [PIP4K2A, phosphatidylinositol, 5-phosphate, 4...
16
        [MAML2, mastermind-like, protein, 2|mam-2|mast...
17
        [PTPRJ, receptor-type, tyrosine-protein, phosp...
[PARD3, partitioning, defective, 3, homolog|CT...
18
19
        [RXFP1, relaxin, receptor, 1|leucine-rich, rep...
20
        [GRB10, growth, factor, receptor-bound, protei...
21
22
23
        [ELMO1, engulfment, and, cell, motility, prote...
24
        [TP53, cellular, tumor, antigen, p53|antigen, ...
25
        [MAGI2, membrane-associated, guanylate, kinase...
        [GSTT1, glutathione, S-transferase, theta-1|GS...
26
        [RIT2, GTP-binding, protein, Rit2|GTP-binding,...
[RSPO2, R-spondin-2|R-spondin, 2, homolog|roof...
27
28
29
        [PYGL, glycogen, phosphorylase, ,, liver, form...
30
        [MTHFR, methylenetetrahydrofolate, reductase|5...
31
        [KCNE4, potassium, voltage-gated, channel, sub...
557
        [SIAE, sialate, O-acetylesterase|H-Lse|cytosol...
558
        [MLXIP, MLX-interacting, protein|Mlx, interact...
559
        [TSGA10, testis-specific, gene, 10, protein|ca...
560
        [DPH1, 2-, (, 3-amino-3-carboxypropyl, ), hist...
        [TPD52L2, tumor, protein, D54|HCCR-binding, pr...
561
562
        [UHMK1, serine/threonine-protein, kinase, Kist...
        [HES5, transcription, factor, HES-5|class, B, ...
563
        [ATF7IP, activating, transcription, factor, 7-...
564
        [RASSF4, ras, association, domain-containing, ...
567
568
        [BLID, BH3-like, motif-containing, cell, death...
        [BCL2L14, apoptosis, facilitator, Bcl-2-like, ...
569
        [FCRL2, Fc, receptor-like, protein, 2|IFGP, fa...
570
571
        [ADGRF1, adhesion, G-protein, coupled, recepto...
572
        [UBE2Q2, ubiquitin-conjugating, enzyme, E2, Q2...
573
        [FCRL1, Fc, receptor-like, protein, 1|IFGP, fa...
        [FHOD3, FH1/FH2, domain-containing, protein, 3...
574
        [TTC12, tetratricopeptide, repeat, protein, 12...
575
        [TAL2, T-cell, acute, lymphocytic, leukemia, p...
[CREBL2, cAMP-responsive, element-binding, pro...
576
578
        [CLECL1, C-type, lectin-like, domain, family, ...
579
                                  [SLX4IP, protein, SLX4IP]
580
        [HMHB1, minor, histocompatibility, protein, HB...
[ZNF296, zinc, finger, protein, 296|zinc, fing...
581
582
                                    [MIR520H, hsa-mir-520h]
583
        [RCSD1, capZ-interacting, protein|RCSD, domain...
584
        [LOC107980440, ABL, proto-oncogene, 1, ,, non-...
585
        [LOC107963955, BCR, p210, Philadelphia, chromo...
586
587
        [LOC107963951, BCR, p190, Philadelphia, chromo...
590
                         [HOXA, @, homeo, box, A, cluster]
             [E2A, single-stranded, DNA-binding, protein]
591
Length: 573, dtype: object
```

In [5]:

```
# remove special characters and numbers
def prepare_text_for_lda(text):
    cleaned = []
    for word in text:
        for element in word:
            cleaned.append(re.sub('[^a-zA-Z+?0-9_]+', '', element)) # remove all special characters and numbers
            cleaned.append(re.sub(r'[^\w]', '', element))
            cleaned = list(dict.fromkeys(cleaned)) # remove duplicates
    return cleaned

df2_tokens = prepare_text_for_lda(df2_tokens)
df2_tokens
```

```
Out[5]:
['IKZF1',
'DNAbinding',
 'protein',
'IkarosCLLassociated',
 'antigen',
'KW6ikaros',
  'family',
 'zinc',
'finger'
  '1lymphoid',
  'transcription',
 'factor',
'LyF1protein',
  'phosphatase',
 '1',
  'regulatory',
  'subunit',
  '92zinc'
  'subfamily',
 '1A',
'Ikaros',
 'ARID5B',
  'ATrich',
  'interactive',
  'domaincontaining',
 '5BARID',
  '5BATrich'
  'domain',
 '5B',
'MRF1like',
  'proteinmodulator',
  'recognition',
 '2',
'MRF2',
'CEBPE',
'CCAATenhancerbinding',
  'epsilonCCAATenhancer',
  'binding',
 'CEBP',
'epsiloncEBP',
  'epsilon',
 'DDC',
'aromaticLaminoacid',
  'decarboxylasedopa',
  'decarboxylase',
'aromatic',
  'Lamino',
 'acid',
  'tetraspanin7CD231',
  'antigenTcell',
 'acute',
'lymphoblastic',
 'leukemia',
'associated',
  '1cell',
  'surface'
  'glycoprotein',
  'A15membrane',
 'component',
'chromosome'
  'Χ',
  'marker',
  '1membrane',
 '1tetraspanin',
'proteintransmembrane',
 'superfamily',
'2btransmembrane',
 'member',
'2transmembrane',
  'A15tspan7',
 'NALT1',
'MIR4674',
 'host',
  'nonprotein',
 'coding',
'MIR4674HG'
  'LINC01573NOTCH1',
 'long',
'noncoding',
  'RNA',
 'in',
  'pediatric',
  'TALLlong',
  'intergenic',
  '1573',
  'CDKN2Á',
```

```
'cyclindependent',
'kinase',
'inhibitor',
'2ACDK4',
'p16INK4alternative',
'reading'
'framecell',
'cycle',
'negative',
'regulator',
'betacyclindependent',
'Acyclindependent',
'2A',
'melanoma',
'p16',
'inhibits',
'CDK4',
'multiple',
'tumor',
'suppressor',
'PDE4B',
'cAMPspecific',
'35cyclic',
'phosphodiesterase',
'4Bduncelike',
'E4phosphodiesterase',
'4B',
'E4',
'dunce',
'homolog',
'Drosophila',
'IL15',
'interleukin15',
'TP63'
'63amplified',
'squamous',
'cell',
'carcinomachronic',
'ulcerative',
'stomatitis',
'proteinkeratinocyte',
'KETtransformationrelated',
'63tumor',
'p53competing',
'GATA3',
'transacting',
'Tcellspecific',
'GATA3GATAbinding',
'3',
'NCOA3',
'nuclear'
'receptor',
'coactivator',
'3CBPinteracting',
'proteinamplified',
'breast',
'cancer',
'proteinclass',
'basic'
'helixloophelix',
'42receptorassociated',
'3steroid',
'3thyroid',
'hormone',
'activator',
'molecule',
'RYR2',
'ryanodine',
'2cardiac',
'muscle',
'receptorcalcium',
'release',
'channelcardiactype',
'receptorkidneytype',
'receptorryanodine',
'cardiac',
'type',
'potassium',
'voltagegated',
'channel',
'A',
'3potassium',
'voltage',
'gated',
'shaker'
'related'
'shakerrelated',
'3type',
```

```
'channelvoltagegated',
'K',
'HuKIIIvoltagegated',
'Kv13voltagegated',
'Kv13',
'PIP4K2A'
'phosphatidylinositol',
 '5phosphate',
'4kinase',
'type2',
'alpha1phosphatidylinositol',
'2alpha1phosphatidylinositol4phosphate',
'kinase1phosphatidylinositol4phosphate5kinasePI',
'5',
'P',
'II',
'alphaPIP4KIIalphaPIP5KIIIPIP5KIIalphaPtdIns',
'P5kinase',
'Β',
'isoformdiphosphoinositide',
'2alphaphosphatidylinositol',
'alphaphosphatidylinositol4phosphate',
'5kinase'
'alphaptdIns',
'C',
'isoformptdIns',
'P4kinase',
'isoform',
'2alphatype'
'phosphatidylinositol4phosphate',
'53',
'MAML2'
'mastermindlike'
'2mam2mastermindlike',
'PTPRJ',
'receptortype'
'tyrosineprotein',
'etaCD148',
'antigenDEP1HPTP',
'etaRPTPJdensityenhanced',
'1human',
'density
'enhanced',
'phosphatase1protein',
 'tyrosine',
'polypeptideproteintyrosine',
'etasusceptibility',
'to',
'colon'
'mouse',
'of',
'PARD3'
'partitioning',
'defective',
'homologCTCL',
'se25PAR3alphaatypical',
'PKC',
'isotypespecific',
'interacting',
'proteinbazookapar3',
'polarity',
'alphapar3'
'homologprotein',
'118',
'RXFP1',
'relaxin',
'1leucinerich'
'repeatcontaining',
'proteincoupled',
'7relaxininsulin',
'like',
'peptide',
'TBXAS1',
'thromboxaneA',
'synthaseTXA',
'synthasecytochrome',
'P450',
'5A1cytochrome',
'polypeptide',
'1platelet',
'cytochrome',
'Vthromboxane',
'synthase',
'platelet',
'GRB10',
'growth',
'receptorbound',
'10GRB10',
```

```
'proteinGRB10',
'adaptor',
'proteininsulin',
'receptorbinding',
'GrbIRmaternally',
'expressed',
'ELMO1',
'engulfment',
'and',
'motility',
'1ced12',
'TP53',
'cellular'
'p53antigen'
 'NYCO13mutant',
'53p53',
'suppressorphosphoprotein',
'p53transformationrelated',
'53tumor',
'supressor',
'p53',
'MAGI2',
'membraneassociated',
'guanylate',
'WW',
'PDZ',
'2activin',
'latrophinlinteracting',
'Amembraneassociated',
'inverted',
'GSTT1',
'glutathione',
'Stransferase',
'theta1GST',
'classtheta1glutathione',
'transferase',
'T11',
'GTPbinding',
'Rit2GTPbinding',
'Roc2Riclike',
'neuronsraslike',
'without',
'CAAX',
'Rspondin2Rspondin',
'homologroof',
'platespecific',
'spondin2',
'PYGL',
'glycogen'
'phosphorylase',
'liver',
'formphosphorylase',
'MTHFR',
'methylenetetrahydrofolate',
'reductase510methylenetetrahydrofolate',
'reductase',
'NADPH',
'NAD',
'H',
'KCNE4',
'4MINKrelated',
'3cardiac',
'accessory',
'4minimum',
'ion',
'channelrelated',
'beta',
'MiRP3potassium',
'4potassium',
'Iskrelated',
'KCNMB2',
'calciumactivated',
'beta2BK',
'2BK',
'beta2MaxiK',
'subunitMaxiK',
'betasubunit',
'2big',
'2charybdotoxin',
'beta2hCG1646471hbeta2k',
'VCA',
'beta2large',
'conductance',
'subunitlargeconductance',
'Ca2+activated',
'Ca2activated',
'K+',
'beta2'
'subunitpotassium',
```

```
'2potassium',
'large',
'2slobeta2',
'SLCO3A1',
'solute',
'carrier'
'organic',
'anion',
'transporter',
'3A1PGE1',
'transporterorganic',
'polypeptiderelated',
'3organic',
'aniontransporting',
'Dsodiumindependent',
'Dsolute',
'21',
'11solute',
'11',
'DGKB'
'diacylglycerol',
'beta90',
'kDa',
'kinaseDAG',
'betadiacylglycerol',
'90kDadiglyceride',
'ABCB1',
'multidrug',
'resistance'
'1ATPbinding',
'cassette',
'MDRTAP',
'1P',
'glycoproteinPglycoprotein',
'1colchicin',
'sensitivitydoxorubicin',
'KRAS',
'GTPase',
'KRasKRas',
'2Kras',
'p21',
'proteinKirsten',
'rat',
'sarcoma',
'viral',
'oncogene'
'homologKirsten',
'protooncogenePR310',
'cKras',
'oncogenecKirascKirstenras',
'proteincellular',
'cKiras2',
'protooncogenecellular',
'transforming',
'protooncogeneoncogene',
'KRAS2transforming',
'p21vKiras2',
'Kirsten',
'MYRIP',
'rab',
'effector',
'MyRIPSlp'
'homologue',
'lacking',
'C2',
'domainsexophilin8myosinVIIa',
'Rabinteracting',
'proteinslp',
'domains',
'csynaptotagminlike',
'domainscsynaptotagminlike',
'KRT85',
'keratin',
'cuticular',
'Hb5hair',
'K212hard',
'5keratin',
'85',
'IIkeratin',
'hair',
'Hb5typeII',
'Kb25',
'myc',
'protooncogene',
'proteinavian',
'myelocytomatosis',
'homologclass',
'39mycrelated',
'translationlocalization',
'factorprotooncogene',
```

```
'cMyctranscription',
'p64vmyc',
'avian',
'homologvmyc',
'BCL2',
'apoptosis',
'Bcl2Bcell'
'CLLlymphoma',
'2protein',
'50',
'ERC',
'2CAZassociated',
'structural',
'proteincytomatrix',
'proteincytomatrix
'p110',
'AIG1',
'androgeninduced',
'NPFFR1',
'neuropeptide',
'FF',
'1Gprotein',
'coupled',
'147RFamiderelated',
'OT7T022neuropeptide',
'ST6GALNAC3',
\verb|'alphaNacetylgalactosaminide'|,
'alpha26sialyltransferase',
'3GalNAc',
'IIIST6',
\verb|'alphaNacetylneuraminy| 123 betagalactosyl 13",
'Nacetylgalactosaminide',
'alpha26sialyltranST6',
'3ST6',
'GalNAc'
'3alphaNacetylgalactosaminide',
'IIIsialyltransferase',
'Nacetyl',
'galactosaminide',
'Csialyltransferase',
'7C',
'230zinc',
'FDZF2',
'INTS10',
'integrator',
'complex',
'10',
'NOTCH1',
'neurogenic',
'locus',
'notch',
'1Notch'
'translocationassociatednotch',
'1translocationassociated',
'TAN1',
'mAAA',
'proteaseinteracting',
'mitochondrial',
'ST8SIA6',
'alpha28sialyltransferase',
'8FST8',
'alphaNacetylneuraminate',
'6sialyltransferase',
'8F',
'alpha2'
'8sialyltransferase',
'sialyltransferase',
'St8Sia',
'VI',
'FLT3',
'FLT3CD135',
'antigenFL',
'cytokine',
'receptorfetal',
'2fmslike',
'3growth',
'IIIstem',
'ABL1',
'nonreceptor',
'kinaseAbelson',
'1bcrcabl',
'proteincabl',
'kinaseprotooncogene',
'cAblprotooncogene',
'ABL1truncated',
'ABL',
'kinasevabl',
'Abelson',
```

```
'murine',
'RUNX1',
'runtrelated',
'1AML1EVI1',
'fusion',
'proteinPEA2alpha',
'BPEBP2alpha',
'BSL33',
'enhancer'
'alpha',
'subunitSL3AKV',
'corebinding',
'subunitacute',
'myeloid',
'proteincorebinding',
'runt',
'2oncogene',
'AMLipolyomavirus',
'enhancerbinding',
'subunitrunt',
'JAK2',
'JAK2JAK2Janus',
'a',
'TYMS'
'thymidylate',
'synthaseTSase',
'KMT2A',
'histonelysine'
'Nmethyltransferase',
'2ACXXCtype',
'7lysine',
'specific',
'methyltransferase',
'2Alysine',
'2Amixed',
'lineage'
'1myeloidlymphoid',
'or',
'mixedlineage',
'trithorax',
'trithoraxlike',
'proteinzinc',
'HRX',
'thiopurine',
'SmethyltransferaseSadenosylLmethionine',
'Smethyltransferase',
'Smethyltransterase
'GCSAML',
'germinal',
'centerassociated',
'signaling',
'motilitylike',
'MARGIGERA'
'MAP1LC3B2',
'microtubuleassociated',
'proteins',
'1A1B',
'light'
'chain',
'2Microtubuleassociated',
'3Blike',
'OR2C3',
'olfactory',
'2C3olfactory',
'2C4olfactory'
'2C5olfactory'
'OR130olfactory',
'4olfactory',
'pseudogene',
'PTEN',
'345trisphosphate',
'3phosphatase',
'dualspecificity',
'PTENMMAC1',
'tensin',
'deleted',
'on',
'10mitochondrial',
'PTENalphamitochondrial',
'alphamutated',
'advanced',
'cancers',
'1phosphatase',
'tensinlike',
'proteinphosphatidylinositol345trisphosphate',
'PTENprotein',
'LMBRD2',
'LMBR1',
'GSTM1',
'Mu',
'1GST',
'HB',
```

```
'classmu',
'1HB',
'4S',
'hydroxyalkyl',
'lyaseglutathione',
'Salkyltransferaseglutathione',
'Saralkyltransferaseglutathione',
'Saryltransferaseglutathione',
'M1',
'ETV6ETS',
'translocation',
'variant',
'6ETSrelated',
'Tel1TEL1',
'oncogeneets',
'6',
'TEL',
'PAX5'
'paired',
'box',
'Pax5Bcell',
'activatorpaired',
'homeotic',
'5paired',
'5transcription',
'PAX',
'AKT1',
'RACalpha',
'serinethreonineprotein',
'kinaseAKT1mPKB',
'alphaRACPKalphaprotein',
'alphaprotooncogene',
'cAktrac',
'alphaserinethreonine',
'kinasevakt',
'thymoma',
'1vakt',
'oncogenelike',
'CDKN2B',
'BCDK',
'inhibitory'
'proteinCDK4B',
 'inhibitorMTS2cyclindependent',
'2B',
'p15',
'kinases',
'proteinmultiple',
'2p14INK4bp14_CDK'
'inhibitorp14_INK4Bp15',
'CDK'
'inhibitorp15INK4bp15 INK4B',
'GSTP1',
'PGST',
'classpiGSTP11deafness',
'Xlinked',
'7epididymis',
'secretory',
'Li',
'22fatty',
'ethyl',
'ester',
'III',
'NRAS'
'NRasNras',
'part',
'4neuroblastoma',
'RAS',
'homologneuroblastoma',
'homologtransforming',
'NRasvras',
'neuroblastoma',
'TCF3',
'E2alphaNOL1TCF3',
'fusionVDR',
'repressorclass'
'21helixloophelix'
'HE47immunoglobulin',
'1kappaE2binding',
'factornegative',
'vitamin',
'D',
'response',
'elementbinding',
'proteintranscription',
'E2A',
'immunoglobulin',
'factors',
'E12E47',
'3transcription',
'ITF1vitamin',
```

```
'receptorinteracting',
'repressor',
'CRLF2',
'receptorlike',
'2ILXRTSLP',
'receptorcytokine',
'CRL2',
'precusorthymic',
'stromal',
'lymphopoietin',
'receptorthymic
'stromalderived',
'FBXW7',
'7Fbox',
'WD',
'repeat',
'containing',
'E3',
'ligaseFbox',
'archipelago',
'Fbox',
'FBW7Fbox'
'FBX30Fbox'
'SEL10archipelagohomolog',
'elegans',
'sel10',
'SLC19A1',
'folate',
'1intestinal',
'1placental',
'transporterreduced',
'1reduced',
'proteinsolute',
'19',
'NQ01'
'dehydrogenase',
'quinone',
'1DTdiaphoraseNAD',
'1NAD',
'Quinone'
'acceptor'
'oxidoreductase',
'menadione',
'oxireductaseazoreductasediaphorase',
'NADHNADPH',
'b5',
'diaphorase4dioxininducible',
'1menadione',
'reductasephylloquinone',
'reductasequinone',
'reductas
'NOTCH3',
'3Notch',
'3notch',
'NR3C1',
'glucocorticoid',
'receptorglucocorticoid',
'1nuclear',
'group',
'hGRB<sup>'</sup>,
'54',
'77',
'93',
'BCR',
'breakpoint',
'cluster',
'region',
'proteinBCR',
'RhoGEF',
'activating'
'proteinBCRFGFR1',
'chimera',
'proteinFGFR1BCR',
'proteinbreakpoint',
'regionrenal',
'carcinoma',
'NYREN26',
'CYP1A1',
'hydrocarbon',
'hydroxylasecytochrome',
'hydroxylasecytochrome',
'Pl450',
'dioxininduciblecytochrome',
'form',
'6cytochrome',
'1cytochrome',
Ί',
'compoundinducible',
'P450Ccytochrome'
'P450P1flavoproteinlinked',
```

```
'monooxygenasexenobiotic',
'monooxygenase',
'TERT',
'telomerase',
'reverse',
'transcriptasetelomerase',
'catalytic',
'subunittelomeraseassociated',
'MTOR',
'12rapamycin',
'2FK506binding',
'complexassociated',
'1FKBPrapamycin',
'proteinFKBP12rapamycin',
'1mammalian',
'target',
'rapamycinmechanistic',
'rapamycin',
'serinethreonine',
'FKBP12',
'1rapamycin',
'FRAP2rapamycin',
'DNMT3A',
'DNA',
'cytosine5',
'3ADNA',
'alphaDNA'
'MTase',
'HsaIIIADNA',
'cytosine',
'3A2',
'TAL1'
'Tcell',
'lymphocytic',
'1Tcell',
'leukemialymphoma',
'5class',
'17stem',
'proteintal1',
'product',
'JAK1',
'PBX1',
'preBcell',
'1homeobox'
'PBX1homeobox',
'PRLpreBcell',
'homeobox',
'WT1',
'Wilms',
'proteinWilms',
'PTPN11',
'11PTP2Cproteintyrosine',
'1Dproteintyrosine',
'2C',
'NFKB1'
'NFkappaB',
'p105',
'subunitDNAbinding',
'KBF1NFkappabetanuclear',
'p50',
'subunitnuclear',
'kappaB',
'kappa',
'Bcells',
'BRAF',
'Braf94',
'Braf'
'proteinBRaf',
'p94',
'BRaf'
'serinethreonineproteinmurine',
'vraf',
'B1protooncogene',
'BRafvraf',
'Bvraf',
'B1',
'resistanceassociated',
'4MRPcMOATrelated',
'AMMPCMMAIrelated',
'ABC',
'transporterbA464I21',
'ATPbinding',
'CFTRMRP',
'canalicular',
'multispecific',
'AFF1',
'AF4FMR2'
'1ALL1fused',
'from',
'proteinmyeloidlymphoid',
'translocated',
```

```
'2preBcell',
'monocytic',
'partner',
'1protooncogene',
'AF4',
'NUDT15'
'nucleotide',
'triphosphate',
'diphosphatase',
'NUDT158oxodGTPase',
'NUDT15mutT',
'2nucleoside'
'diphosphatelinked',
'another',
'moiety',
'hydrolase',
'15nudix',
'nucleoside',
'diphosphate',
'linked',
'motif',
'15probable',
'78dihydro8oxoguanine',
'triphosphatase',
'NUDT15probable',
'8oxodGTP',
'CXCR4',
'CXC',
'chemokine',
'4CD184',
'antigenLPSassociated',
'3SDF1',
'receptorchemokine',
'4chemokine',
'receptorfusinleukocytederived',
'seven',
'transmembrane'
'receptorlipopolysaccharideassociated',
'3neuropeptide',
'Y3neuropeptide',
'Y3',
'receptorseven',
'helix',
'receptorseventransmembranesegment',
'spleenstromal',
'cellderived',
'MDM2',
'ubiquitinprotein',
'ligase',
'Mdm2MDM2'
'ligaseMDM2'
'ligaseMdm2',
'homologMdm2',
'transformed',
'3T3',
'double',
'minute',
'proteindouble',
'human',
'p53binding',
'proteinoncoprotein',
'Mdm2',
'IGH',
'diversity',
'heavy',
'chainsJ'
'joining',
'chainsimmunglobulin',
'variable',
'regionimmunoglobulin',
'clusterimmunoglobulin',
'groupimmunoglobulin',
'locusimmunoglobulin',
'BCL2L11',
'bcl2like',
'11BCL2like'
'facilitator',
'bcl2',
'mediator'
'deathbcl2',
'Bimbcl2related',
...]
```

```
In [6]:
```

```
# define stopwords (genome specific)
  stopword_list = stopwords.words('english')
 print(stopword_list)
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'd", "you'd", 'yours', 'yourself', imself', 'she's", 'her', 'herself', 'it', "it's", 'its!, 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what' hat', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'three', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'the 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'to', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "could esn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn't", 'needn', n', "shouldn't", 'wasn', "wasn't", 'weren't", 'won', "won't", 'wouldn', "wouldn't", 'sapiens', 'homo', '9606', 'single', 'minus' nd', 'nc_001460.1', 'receptor', 'factor', 'subunit', 'kinase', 'class', 'homolog', 'member', 'alpha', 'oncogene', 'transcription', 'helix
 In [7]:
  # Build the bigram and trigram models
  bigram = gensim.models.Phrases(df2_tokens, min_count=5, threshold=100) # higher threshold fewer phrases.
  trigram = gensim.models.Phrases(bigram[df2_tokens], threshold=100)
  # Faster way to get a sentence clubbed as a trigram/bigram
  bigram_mod = gensim.models.phrases.Phraser(bigram)
  trigram_mod = gensim.models.phrases.Phraser(trigram)
  # Define functions for stopwords, bigrams, trigrams and Lemmatization
  def remove_stopwords(texts):
         return [w for w in texts if not w in stopword_list]
  def make_bigrams(texts):
         return [bigram_mod[doc] for doc in texts]
  def make trigrams(texts):
         return [trigram_mod[doc] for doc in texts]
```

Call the above declared methods

In [8]:

```
data_words_nostops = remove_stopwords(df2_tokens)

data_words_bigrams = make_bigrams(data_words_nostops)

data_words_trigrams = make_trigrams(data_words_bigrams)

print(data_words_bigrams)
```

```
'D',
'e',
                    'r', 'o', 'p', 'h', 'i', 'l'], ['p', 'e', 'p', 't', 'i', 'd', 'k', 'e'], ['p', 'r', 'o', 't', 'e', 'i', 'n', 'c', 'h', 'e', 'o', 'c', 'y', 't', 'e'], ['1', 'i', 'n', 't', 'e', 'r', 'l', 'o', 'm', 'a', 'd'], ['p', 'r', 'i', 'v', 'e', 'd'], ['p', 'r', 'l', 'e', 'm', 'o', 'n', 'o', 'c', 't', 'e', 'd', 'e', 'r', 'i', 'v', 'e', 'd'], ['n', 'e', 'u',
                                                                                                                                                                   'a,
'n', 'e',
'n'], ['8', '1
'n', 'l', 'y',
'i', 'i',
                                                                                                                      'e',
'l',
                                                                                                                                                                                        '8', 'e', ...
'l', 'u', 'n', 'g'],
'', 'm', 'p', 'h', 'o'
'e', 'd'], ['
                                                                                                                                                          'i',
                                                                                                                                                          'i',
   'i',
           'n',
                   'o', 'm',
                                                                                                                                'o',
                                                                                                                                         't',
                                                                                                                                                 'e',
                                                                                                                                         't',
                                                                                                                                                 'e', 'd',
'o', 'p',
           'd'],
                                                                                                                                                                           'r',
'i',
                                                                                                                                't',
                                                                                                                                                                   'h',
                                                                                                                                                                                    '1',
                                                                                                                                                                                                      'c',
                                                                                                                                                                                                              't',
                                                                                                                                         'r',
                                                                                                                                                                                             'a',
```

Subject to the control of the contro ''o', 'n', 'e', 'c', 't', 'i a', 'c', 't', 'i', 'v', 'a', ', 'y', 't', 'e'], ['C', 'D' interest (1986) | The second of the second o

State Segments of the content of the , '1',

```
'n', 'u
'1',
'1
```

```
['h', 'o', 'm
d', 'a', 's', 'e',
'o', 'p', 'e', '!
'', 'p', 'e',
'v', 'e',
```

subsetS8868coverthminks all genomes v04 sympars and set in the set of the set 'e', 's', 'y', 'l', 'g', 'a', 'l', 'a', 'c', 't', 'o', 's', 'y', 'l', 'g', 'a', 'l', 'a', 'c', 't', 'o', 's', 'y', 'l', 't', 'o', 's', 'y', 'l', 't', 'o', 'n', 'o', 's', 'y', 'l', 't', 'U', 'D', 'P', 'G', 'l', 'c', 'U', 'A'], ['b', 'e', 't', 'e', 'g', 'l', 'u', 'c', 'u', 'r', 'o', 'n', 'o', 's', 'y', 'n', 'A', 'L', 'L', 'l', 'f', 'u', 's', 'e', 'd'], ['4', 'm',

South See The Company of the Company

'i',
', 'P', 'A
'M', 'P', 'a',
'o'], ['2', 'm',
'c', 'R', 'l',
'C', 'A', 'e'], ['I', B, n], L, ', 'r', 'l', 'l', 'l', 'k', 'e'], ['I', 'M', 'P', A, 'a', 't', 'a', 's', 'e'], ['2', 'I', 'M', 'P'], ['2', 'I', 'M', 'P'], ['2', 'I', 'M', 'P'], ['2', 'I', 'M', 'P'], 'a'], ['F', 'C', 'R', 'L', '5'], ['5', 'f', 'c'], ['5', 'f', 'c', 'R', 'l', 'o', 'b', 'u', 'l', 'i', 'n'], ['I', 'R, 'T', 'A', '2'], ['C', 'C', 'A'], ['C', 'B', 'L', 'I', 'F'], ['i', 'n', 't', 'r', 'a', 'a', 'a'], ['C', 'B', 't', 'e', 'r', 'o', 'l', 'b', 'a', 'n', 'd', 'i', 'n', 'd'], ['1', 'a', 's', 't', 'e', 'r', 'o', 'l', 'b', 'a', 'a', 's'], ['', 'o', 'l'], ['1', 'o', 'l'], ['1', 'o', 'l'], ['1', 'o', 'l', 'o' n', 's', 'c', 'r', '
k', 'e'], ['I', 'M',
'P'], ['2', 'I', '
'1'], ['m', 'y', 'o
'c'], ['5', 'f', 'c
, 'T', 'A', '2'], [' ', к 'М', . '1'], ['F', ', 'o', 'b', 'c'], [' 'b', 1,
, ['G', 'P', 'R,
'A', 'P'], ['7', 'M'
'd'], ['L', 'Y', 'L'
't', 'o', 'p', 'o', '
, 'X', '2', '2'], ['N'
' 'n', '2'], ['L', ' 'u', 'm', 'u', 'l' 'a', 's', 'e', '7', 'a', 't', 'i', 'o', 'm', 'i', 't', 'o', 'p', ',
K', 'A
'i', 'n',
'o', 's',
', 'i',
'c', 'h', 'o', 'b', 'l', 'a', 's', 't', '
'2'], ['B', 'h', 'o', 'm', 'e', 'o',
, 'W', 'i', 'l', 'l', 'i', 'a', 'm',
'n', 'W', 'i', 'l', 'l', 'i', 'a', ', 'o',
', 's', 's',
'S'], ['2', 'a',
['0', 'a', 'c', 'e',
'c'], ['9', '0', 'a',
'', 's', 'p', 'e',
'6', 't' 't',
's', 'o
''. 'a', 'n
''. 'e', 't
''. 'c',
''. 'c', 'e',
'o', 'c
'. 'n', 'k
'. 't', 'y
'. 'e', 'n', '1', 'a', 'i', 'n', 'e', 'n', 'T', c, 'n', 'g'], ['S', 'I', 's', 's', 's', 's', 'c', 'T', 'o', 'n', 'y', g',
't', 'o',
', 'l', 'o', 'g',
's', 'e'], ['M', 'L',
't', 'e', 'r', 'a', 'c'
'o', 'A'], ['T',
't', 'e', ', 'm', 'a', 'r ',
'o',
's',
't 'a', 'c', 'l'], ['M', 'd', 'd', 'd', 'd', 'd', 'd', 'a', 'o', 'n', 'n', 'c', 'a', 'n', 'c', 'p', 'e', 'c', 'i', 'f', 's', 't', 'i', 'd', 'i', 'k', 'e'], ['1', 'S', 'a' 's'], ['N', '1[']], [[']h', n', 'i', 'n', 'C,
'u', '1
'. 'm', 'u
'. '1',
'a' i', 'm ,
'G'], ['1', ' '2'], ['q', 'R'], ['h', Ō',

Create a dictionary (from bigrams) and form a corpus

```
In [9]:
# Create Dictionary
id2word = corpora.Dictionary(data_words_bigrams)
# Create Corpus
texts = data_words_bigrams
# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]
# View
print(corpus[:1])
[[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1)]]
In [10]:
id2word[0]
Out[10]:
'1'
In [11]:
# Human readable format of corpus (term-frequency)
[[(id2word[id], freq) for id, freq in cp] for cp in corpus[:1]]
Out[11]:
[[('1', 1), ('F', 1), ('I', 1), ('K', 1), ('Z', 1)]]
Build LDA Model (using corpus and dictionary)

    5 topics

 • 10 passes
In [12]:
```

In [13]:

```
# Print the Keyword in the 10 topics
print(lda_model.print_topics())
doc_lda = lda_model[corpus]

[(0, '0.248*"i" + 0.165*"n" + 0.137*"k" + 0.097*"b" + 0.095*"2" + 0.083*"u" + 0.055*"v" + 0.048*"g" + 0.036*"d" + 0.029*"q"'), (1, '0.226
*"L" + 0.076*"T" + 0.065*"3" + 0.063*"S" + 0.048*"H" + 0.044*"5" + 0.038*"0"'), (2, '0.117*"e" + 0.091*"i" + 0.086*"a" + 0.082*"t" + 0.082
2*"l" + 0.061*"S" + 0.052*"c"'), (3, '0.177*"C" + 0.155*"D" + 0.148*"2" + 0.098*"4" + 0.098*"B" + 0.069*"9" + 0.068*"A" + 0.064*"K" + 0.0
+ 0.148*"I" + 0.146*"M" + 0.128*"F" + 0.085*"E" + 0.085*"2" + 0.084*"G" + 0.065*"7" + 0.029*"H" + 0.019*"Y"'')]
```

Calculate perplexity (to measure the quality of model) and calculate coherence score (more precise than perplexity)

```
In [14]:
```

```
# Compute Perplexity
print('\nPerplexity: ', lda_model.log_perplexity(corpus)) # the lower the better.

# Compute Coherence Score
coherence_model_lda = CoherenceModel(model=lda_model, texts=data_words_trigrams, dictionary=id2word, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score: ', coherence_lda)
Perplexity: -3.411499894616546
```

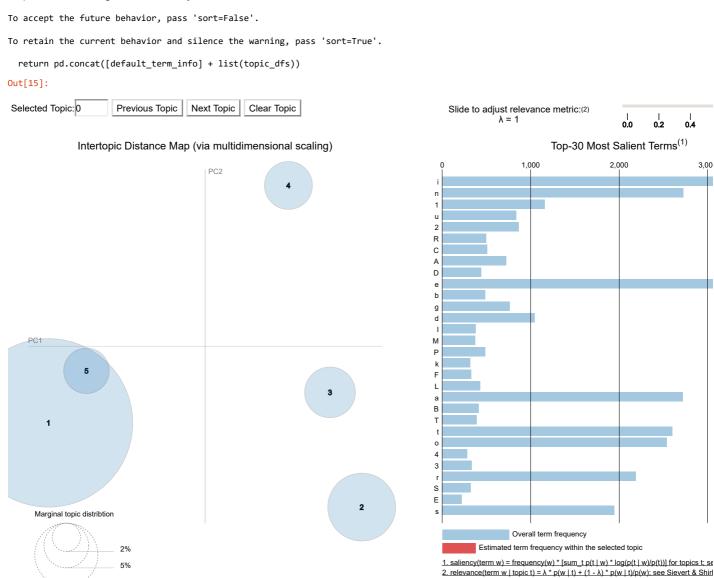
Coherence Score: 0.46729248820258135

Visualize topics

In [15]:

```
pyLDAvis.enable_notebook()
vis = pyLDAvis.gensim.prepare(lda_model, corpus, id2word)
vis
```

C:\Users\jacky\Anaconda3\lib\site-packages\pyLDAvis_prepare.py:257: FutureWarning: Sorting because non-concatenation axis is not aligned of pandas will change to not sort by default.



10%