Information Retrieval
Assignment-1

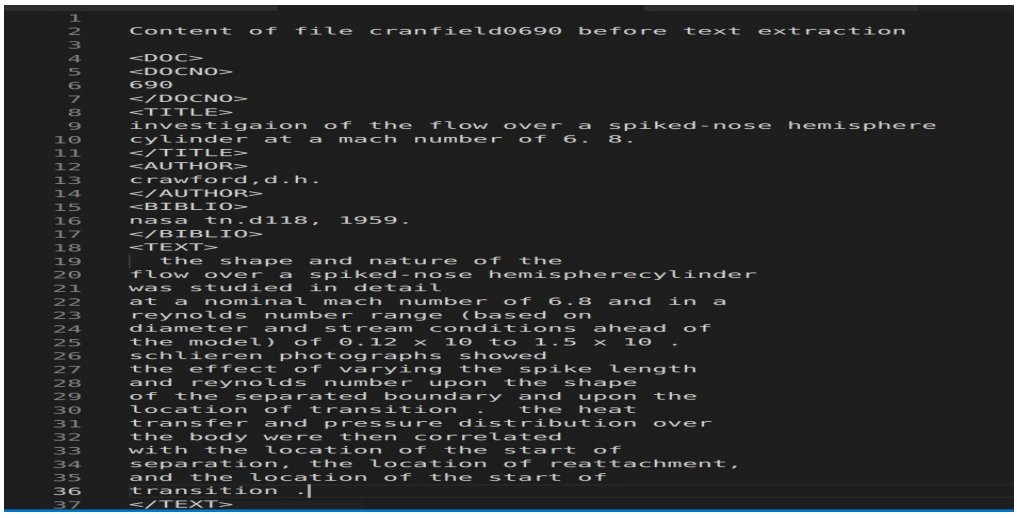
Shubham Agrawal (MT22124)
Amrita Aash (MT22011)
Ayush Agarwal (MT22095)

Question 1:

Methodology-

- ➔ We extracted the text between TITLE and TEXT tags using the following regular expressions:
  - ◆ ((?<=\<TITLE\>)(.|\n)*(?=\<\/TITLE\>))
  - ◆ ((?<=\<TEXT\>)(.|\n)*(?=\<\/TEXT\>))
- ➔ Then we concatenated the texts and overwrote it on the corresponding file.
- ➔ Then for each of the pre-processing steps we created a function to show the contents of the file before & after pre-processing.
- ➔ The following result section shows each of the pre-processing steps.

Result-



```
the body were then correlated
with the location of the start of
separation, the location of reattachment,
and the location of the start of
transition .
</TEXT>
</DOC>


Content of file cranfield0690 after text extraction

investigaion of the flow over a spiked-nose hemisphere cylinder at a mach number of 6. 8. the shape and nature of the flow over a spiked-nose hemisp

------------------------------------------------------------
```

```
# Q2) ii) 1
lowercase_df = pd.DataFrame(columns = ['Doc Id','Before Lowercase','After Lowercase'])
toLowerCase(path2,lowercase_df)
lowercase_df.set_index('Doc Id',inplace = True)
lowercase_df.head()
```

[15] ✓ 3.4s Python

Doc Id	Before Lowercase	After Lowercase
cranfield0690	investigaion of the flow over a spiked-nose hemisphere cylinder at a mach number of 6. 8. the shape and nature of the flow over a spiked-nose hemispherecylinder was studied in detail at a nominal mach number of 6.8 and in a reynolds number range (based on diameter and stream conditions ahead of the model) of 0.12 x 10 to 1.5 x 10 . schlieren photographs showed the effect of varying the spike length and reynolds number upon the shape of the separated boundary and upon the location of transition . the heat transfer and pressure distribution over the body were then correlated with the location of the start of separation, the location of reattachment, and the location of the start of transition .	investigaion of the flow over a spiked-nose hemisphere cylinder at a mach number of 6. 8. the shape and nature of the flow over a spiked-nose hemispherecylinder was studied in detail at a nominal mach number of 6.8 and in a reynolds number range (based on diameter and stream conditions ahead of the model) of 0.12 x 10 to 1.5 x 10 . schlieren photographs showed the effect of varying the spike length and reynolds number upon the shape of the separated boundary and upon the location of transition . the heat transfer and pressure distribution over the body were then correlated with the location of the start of separation, the location of reattachment, and the location of the start of transition .
cranfield1184	three dimensional effects in viscous wakes . three-dimensionality in wakelike or jetlike free mixing may stem from initial geometric configurations, nonuniformities in flow variables over a cross section, or boundary conditions along the flow . these may be generated by bodies at angle of attack, nonaxisymmetric bodies, mixing of nonaxisymmetric jets with an outer flow, finite wings, or more artificial means . this paper is devoted to studies bearing on such configurations . the first section deals with the general mathematical model, in which the boundary layer approximations are used, and with methods of solution . laminar and turbulent flow, compressibility, unsteadiness, and streamwise pressure gradients are admitted initially . the flux forms of the equations are given . algebraic integrals of the energy equations and the diffusion (frozenflow) equations are obtained . a simplification of the convective terms, roughly corresponding to the oseen approximation, is used in the asymptotic downstream region . the second section contains explicit solutions for specific configurations, in particular for flows whose initial isovels are of elliptic shape . these flows may be wakelike or jetlike . compressibility is admitted,. however, the flows must have uniform pressure and must be steady . the final section deals with interpretation and evaluation of the results .	three dimensional effects in viscous wakes . three-dimensionality in wakelike or jetlike free mixing may stem from initial geometric configurations, nonuniformities in flow variables over a cross section, or boundary conditions along the flow . these may be generated by bodies at angle of attack, nonaxisymmetric bodies, mixing of nonaxisymmetric jets with an outer flow, finite wings, or more artificial means . this paper is devoted to studies bearing on such configurations . the first section deals with the general mathematical model, in which the boundary layer approximations are used, and with methods of solution . laminar and turbulent flow, compressibility, unsteadiness, and streamwise pressure gradients are admitted initially . the flux forms of the equations are given . algebraic integrals of the energy equations and the diffusion (frozenflow) equations are obtained . a simplification of the convective terms, roughly corresponding to the oseen approximation, is used in the asymptotic downstream region . the second section contains explicit solutions for specific configurations, in particular for flows whose initial isovels are of elliptic shape . these flows may be wakelike or jetlike . compressibility is admitted,. however, the flows must have uniform pressure and must be steady . the final section deals with interpretation and evaluation of the results .

```
# Q2) ii) 2
tokenize_df = pd.DataFrame(columns = ['Doc Id','Before Tokenization','After Tokenization'])
tokenizer(path2,tokenize_df,lowercase_df)
tokenize_df.set_index('Doc Id',inplace = True)
tokenize_df.head()
```

[17] ✓ 5.2s Python

Doc Id	Before Tokenization	After Tokenization
cranfield0690	investigaion of the flow over a spiked-nose hemisphere cylinder at a mach number of 6. 8. the shape and nature of the flow over a spiked-nose hemispherecylinder was studied in detail at a nominal mach number of 6.8 and in a reynolds number range (based on diameter and stream conditions ahead of the model) of 0.12 x 10 to 1.5 x 10 . schlieren photographs showed the effect of varying the spike length and reynolds number upon the shape of the separated boundary and upon the location of transition . the heat transfer and pressure distribution over the body were then correlated with the location of the start of separation, the location of reattachment, and the location of the start of transition .	['investigaion', 'of', 'the', 'flow', 'over', 'a', 'spiked-nose', 'hemisphere', 'cylinder', 'at', 'a', 'mach', 'number', 'of', '6', '.', '8', '.', 'the', 'shape', 'and', 'nature', 'of', 'the', 'flow', 'over', 'a', 'spiked-nose', 'hemispherecylinder', 'was', 'studied', 'in', 'detail', 'at', 'a', 'nominal', 'mach', 'number', 'of', '6.8', 'and', 'in', 'a', 'reynolds', 'number', 'range', '(', 'based', 'on', 'diameter', 'and', 'stream', 'conditions', 'ahead', 'of', 'the', 'model', ')', 'of', '0.12', 'x', '10', 'to', '1.5', 'x', '10', '.', 'schlieren', 'photographs', 'showed', 'the', 'effect', 'of', 'varying', 'the', 'spike', 'length', 'and', 'reynolds', 'number', 'upon', 'the', 'shape', 'of', 'the', 'separated', 'boundary', 'and', 'upon', 'the', 'location', 'of', 'transition', '.', 'the', 'heat', 'transfer', 'and', 'pressure', 'distribution', 'over', 'the', 'body', 'were', 'then', 'correlated', 'with', 'the', 'location', 'of', 'the', 'start', 'of', 'separation', ',', 'the', 'location', 'of', 'reattachment', ',', 'and', 'the', 'location', 'of', 'the', 'start', 'of', 'transition', '.']
cranfield1184	three dimensional effects in viscous wakes . three-dimensionality in wakelike or jetlike free mixing may stem from initial geometric configurations, nonuniformities in flow variables over a cross section, or boundary conditions along the flow . these may be generated by bodies at angle of attack, nonaxisymmetric bodies, mixing of nonaxisymmetric jets with an outer flow, finite wings, or more artificial means . this paper is devoted to studies bearing on such configurations . the first section deals with the general mathematical model, in which the boundary layer approximations are used, and with methods of solution . laminar and turbulent flow, compressibility, unsteadiness, and streamwise pressure gradients are admitted initially . the flux forms of the equations are given . algebraic integrals of the energy equations and the diffusion (frozenflow) equations are obtained . a simplification of the convective terms, roughly corresponding to the oseen approximation, is used in the asymptotic downstream region . the second section contains explicit solutions for specific	['three', 'dimensional', 'effects', 'in', 'viscous', 'wakes', '.', 'three-dimensionality', 'in', 'wakelike', 'or', 'jetlike', 'free', 'mixing', 'may', 'stem', 'from', 'initial', 'geometric', 'configurations', ',', 'nonuniformities', 'in', 'flow', 'variables', 'over', 'a', 'cross', 'section', ',', 'or', 'boundary', 'conditions', 'along', 'the', 'flow', '.', 'these', 'may', 'be', 'generated', 'by', 'bodies', 'at', 'angle', 'of', 'attack', ',', 'nonaxisymmetric', 'bodies', ',', 'mixing', 'of', 'nonaxisymmetric', 'jets', 'with', 'an', 'outer', 'flow', ',', 'finite', 'wings', ',', 'or', 'more', 'artificial', 'means', ',', 'this', 'paper', 'is', 'devoted', 'to', 'studies', 'bearing', 'on', 'such', 'configurations', '.', 'the', 'first', 'section', 'deals', 'with', 'the', 'general', 'mathematical', 'model', ',', 'in', 'which', 'the', 'boundary', 'layer', 'approximations', 'are', 'used', ',', 'and', 'with', 'methods', 'of', 'solution', '.', 'laminar', 'and', 'turbulent', 'flow', ',', 'compressibility', ',', 'unsteadiness', ',', 'and', 'streamwise', 'pressure', 'gradients', 'are', 'admitted', 'initially', '.', 'the', 'flux', 'forms', 'of', 'the', 'equations', 'are', 'given', '.', 'algebraic', 'integrals', 'of', 'the', 'energy', 'equations', 'and', 'the', 'diffusion', '(', 'frozenflow', ')', 'equations', 'are', 'obtained', '.', 'a', 'simplification', 'of', 'the', 'convective', 'terms', ',', 'roughly', 'corresponding', 'to', 'the', 'oseen', 'approximation', ',', 'is', 'used', 'in', 'the', 'asymptotic', 'downstream', 'region', '.', 'the', 'second', 'section', 'contains', 'explicit', 'solutions', 'for', 'specific', 'configurations', ',',

```
# Q2) ii) 3
stopwords_df = pd.DataFrame(columns = ['Doc Id','Before Stopwords Removal','After Stopwords Removal'])
removeStopWords(path2,stopwords_df,tokenize_df)
stopwords_df.set_index('Doc Id',inplace = True)
stopwords_df.head()
```

[19] ✓ 25.7s Python

Doc Id	Before Stopwords Removal	After Stopwords Removal
cranfield0690	['investigaion', 'of', 'the', 'flow', 'over', 'a', 'spiked-nose', 'hemisphere', 'cylinder', 'at', 'a', 'mach', 'number', 'of', '6.', ',', '8.', 'the', 'shape', 'and', 'nature', 'of', 'the', 'flow', 'over', 'a', 'spiked-nose', 'hemispherecylinder', 'was', 'studied', 'in', 'detail', 'at', 'a', 'nominal', 'mach', 'number', 'of', '6.8', 'and', 'in', 'a', 'reynolds', 'number', 'range', '(', 'based', 'on', 'diameter', 'and', 'stream', 'conditions', 'ahead', 'of', 'the', 'model', ')', 'of', '0.12', 'x', '10', 'to', '1.5', 'x', '10', ',', 'schlieren', 'photographs', 'showed', 'the', 'effect', 'of', 'varying', 'the', 'spike', 'length', 'and', 'reynolds', 'number', 'upon', 'the', 'shape', 'of', 'the', 'separated', 'boundary', 'and', 'upon', 'the', 'location', 'of', 'transition', ',', 'the', 'heat', 'transfer', 'and', 'pressure', 'distribution', 'over', 'the', 'body', 'were', 'then', 'correlated', 'with', 'the', 'location', 'of', 'the', 'start', 'of', 'separation', ',', 'the', 'location', 'of', 'reattachment', ',', 'and', 'the', 'location', 'of', 'the', 'start', 'of', 'transition', '.']	['investigaion', 'flow', 'spiked-nose', 'hemisphere', 'cylinder', 'mach', 'number', '6', ',', '8.', 'shape', 'nature', 'flow', 'spiked-nose', 'hemispherecylinder', 'studied', 'detail', 'nominal', 'mach', 'number', '6.8', 'reynolds', 'number', 'range', '(', 'based', 'diameter', 'stream', 'conditions', 'ahead', 'model', ')', '0.12', 'x', '10', '1.5', 'x', '10', ',', 'schlieren', 'photographs', 'showed', 'effect', 'varying', 'spike', 'length', 'reynolds', 'number', 'upon', 'shape', 'separated', 'boundary', 'upon', 'location', 'transition', ',', 'heat', 'transfer', 'pressure', 'distribution', 'body', 'correlated', 'location', 'start', 'separation', ',', 'location', 'reattachment', ',', 'location', 'start', 'transition', '.']
cranfield1184	['three', 'dimensional', 'effects', 'in', 'viscous', 'wakes', ',', 'three-dimensionality', 'in', 'wakelike', 'or', 'jetlike', 'free', 'mixing', 'may', 'stem', 'from', 'initial', 'geometric', 'configurations', ',', 'nonuniformities', 'in', 'flow', 'variables', 'over', 'a', 'cross', 'section', 'or', 'boundary', 'conditions', 'along', 'the', 'flow', ',', 'these', 'may', 'be', 'generated', 'by', 'bodies', 'at', 'angle', 'of', 'attack', ',', 'nonaxisymmetric', 'bodies', ',', 'mixing', 'of', 'nonaxisymmetric', 'jets', 'with', 'an', 'outer', 'flow', ',', 'finite', 'wings', ',', 'or', 'more', 'artificial', 'means', ',', 'this', 'paper', 'is', 'devoted', 'to', 'studies', 'bearing', 'on', 'such', 'configurations', ',', 'the', 'first', 'section', 'deals', 'with', 'the', 'general', 'mathematical', 'model', ',', 'in', 'which', 'the', 'boundary', 'layer', 'approximations', 'are', 'used', ',', 'and', 'with', 'methods', 'of', 'solution', ',', 'laminar', 'and', 'turbulent', 'flow', ',', 'compressibility', ',', 'unsteadiness', ',', 'and', 'streamwise', 'pressure', 'gradients', 'are', 'admitted', 'initially', ',', 'the', 'flux', 'forms', 'of', 'the', 'equations', 'are', 'given', ',', 'algebraic', 'integrals', 'of', 'the', 'energy', 'equations', 'and', 'the', 'diffusion', '(', 'frozenflow', ')', 'equations', 'are', 'obtained', ',', 'a', 'simplification', 'of', 'the', 'convective', 'terms', ',', 'roughly', 'corresponding', 'to', 'the', 'oseen', 'approximation', ',', 'is', 'used', 'in', 'the', 'asymptotic', 'downstream', 'region', ',', 'the', 'second', 'section', 'contains', 'explicit', 'solutions', 'for', 'specific', 'configurations', ',', 'in', 'particular', 'for', 'flows', 'whose', 'initial', 'isovels', 'are', 'of', 'elliptic', 'shape', ',', 'these', 'flows', 'may', 'wakelike', 'jetlike', 'compressibility', 'admitted', 'however', 'flows', 'must', 'uniform', 'pressure', 'must', 'steady', 'final', 'section', 'deals', 'with', 'interpretation', 'evaluation', 'results', '.']	['three', 'dimensional', 'effects', 'viscous', 'wakes', ',', 'three-dimensionality', 'wakelike', 'jetlike', 'free', 'mixing', 'may', 'stem', 'initial', 'geometric', 'configurations', ',', 'nonuniformities', 'flow', 'variables', 'cross', 'section', ',', 'boundary', 'conditions', 'along', 'flow', ',', 'may', 'generated', 'bodies', 'angle', 'attack', ',', 'nonaxisymmetric', 'bodies', ',', 'mixing', 'nonaxisymmetric', 'jets', 'outer', 'flow', ',', 'finite', 'wings', ',', 'artificial', 'means', ',', 'paper', 'devoted', 'studies', 'bearing', 'configurations', ',', 'first', 'section', 'deals', 'general', 'mathematical', 'model', ',', 'boundary', 'layer', 'approximations', 'used', ',', 'methods', 'solution', ',', 'laminar', 'turbulent', 'flow', ',', 'compressibility', ',', 'unsteadiness', ',', 'streamwise', 'pressure', 'gradients', 'admitted', 'initially', ',', 'flux', 'forms', 'equations', 'given', ',', 'algebraic', 'integrals', 'energy', 'equations', 'diffusion', '(', 'frozenflow', ')', 'equations', 'obtained', ',', 'simplification', 'convective', 'terms', ',', 'roughly', 'corresponding', 'oseen', 'approximation', ',', 'used', 'asymptotic', 'downstream', 'region', ',', 'second', 'section', 'contains', 'explicit', 'solutions', 'specific', 'configurations', ',', 'particular', 'flows', 'whose', 'initial', 'isovels', 'elliptic', 'shape', ',', 'flows', 'may', 'wakelike', 'jetlike', 'compressibility', 'admitted', 'however', 'flows', 'must', 'uniform', 'pressure', 'must', 'steady', 'final', 'section', 'deals', 'interpretation', 'evaluation', 'results']

```
# Q2) ii) 4
pucntuation_df = pd.DataFrame(columns = ['Doc Id','Before Punctuation Removal','After Punctuation Removal'])
removePunctuation(path2,pucntuation_df,stopwords_df)
pucntuation_df.set_index('Doc Id',inplace = True)
pucntuation_df.head()
```

[21] ✓ 3.6s Python

Doc Id	Before Punctuation Removal	After Punctuation Removal
cranfield0690	['investigaion', 'flow', 'spiked-nose', 'hemisphere', 'cylinder', 'mach', 'number', '6', ',', '8.', 'shape', 'nature', 'flow', 'spiked-nose', 'hemispherecylinder', 'studied', 'detail', 'nominal', 'mach', 'number', '6.8', 'reynolds', 'number', 'range', '(', 'based', 'diameter', 'stream', 'conditions', 'ahead', 'model', ')', '0.12', 'x', '10', '1.5', 'x', '10', ',', 'schlieren', 'photographs', 'showed', 'effect', 'varying', 'spike', 'length', 'reynolds', 'number', 'upon', 'shape', 'separated', 'boundary', 'upon', 'location', 'transition', ',', 'heat', 'transfer', 'pressure', 'distribution', 'body', 'correlated', 'location', 'start', 'separation', ',', 'location', 'reattachment', ',', 'location', 'start', 'transition', '.']	['investigaion', 'flow', 'spiked-nose', 'hemisphere', 'cylinder', 'mach', 'number', '6', '8.', 'shape', 'nature', 'flow', 'spiked-nose', 'hemispherecylinder', 'studied', 'detail', 'nominal', 'mach', 'number', '6.8', 'reynolds', 'number', 'range', 'based', 'diameter', 'stream', 'conditions', 'ahead', 'model', '0.12', 'x', '10', '1.5', 'x', '10', 'schlieren', 'photographs', 'showed', 'effect', 'varying', 'spike', 'length', 'reynolds', 'number', 'upon', 'shape', 'separated', 'boundary', 'upon', 'location', 'transition', 'heat', 'transfer', 'pressure', 'distribution', 'body', 'correlated', 'location', 'start', 'separation', 'location', 'reattachment', 'location', 'start', 'transition']
cranfield1184	['three', 'dimensional', 'effects', 'viscous', 'wakes', ',', 'three-dimensionality', 'wakelike', 'jetlike', 'free', 'mixing', 'may', 'stem', 'initial', 'geometric', 'configurations', ',', 'nonuniformities', 'flow', 'variables', 'cross', 'section', ',', 'boundary', 'conditions', 'along', 'flow', ',', 'may', 'generated', 'bodies', 'angle', 'attack', ',', 'nonaxisymmetric', 'bodies', ',', 'mixing', 'nonaxisymmetric', 'jets', 'outer', 'flow', ',', 'finite', 'wings', ',', 'artificial', 'means', ',', 'paper', 'devoted', 'studies', 'bearing', 'configurations', ',', 'first', 'section', 'deals', 'general', 'mathematical', 'model', ',', 'boundary', 'layer', 'approximations', 'used', ',', 'methods', 'solution', ',', 'laminar', 'turbulent', 'flow', ',', 'compressibility', ',', 'unsteadiness', ',', 'streamwise', 'pressure', 'gradients', 'admitted', 'initially', ',', 'flux', 'forms', 'equations', 'given', ',', 'algebraic', 'integrals', 'energy', 'equations', 'diffusion', '(', 'frozenflow', ')', 'equations', 'obtained', ',', 'simplification', 'convective', 'terms', ',', 'roughly', 'corresponding', 'oseen', 'approximation', ',', 'used', 'asymptotic', 'downstream', 'region', ',', 'second', 'section', 'contains', 'explicit', 'solutions', 'specific', 'configurations', ',', 'particular', 'flows', 'whose', 'initial', 'isovels', 'elliptic', 'shape', ',', 'flows', 'may', 'wakelike', 'jetlike', 'compressibility', 'admitted', 'however', 'flows', 'must', 'uniform', 'pressure', 'must', 'steady', 'final', 'section', 'deals', 'interpretation', 'evaluation', 'results', '.']	['three', 'dimensional', 'effects', 'viscous', 'wakes', 'three-dimensionality', 'wakelike', 'jetlike', 'free', 'mixing', 'may', 'stem', 'initial', 'geometric', 'configurations', 'nonuniformities', 'flow', 'variables', 'cross', 'section', 'boundary', 'conditions', 'along', 'flow', 'may', 'generated', 'bodies', 'angle', 'attack', 'nonaxisymmetric', 'bodies', 'mixing', 'nonaxisymmetric', 'jets', 'outer', 'flow', 'finite', 'wings', 'artificial', 'means', 'paper', 'devoted', 'studies', 'bearing', 'configurations', 'first', 'section', 'deals', 'general', 'mathematical', 'model', 'boundary', 'layer', 'approximations', 'used', 'methods', 'solution', 'laminar', 'turbulent', 'flow', 'compressibility', 'unsteadiness', 'streamwise', 'pressure', 'gradients', 'admitted', 'initially', 'flux', 'forms', 'equations', 'given', 'algebraic', 'integrals', 'energy', 'equations', 'diffusion', 'frozenflow', 'equations', 'obtained', 'simplification', 'convective', 'terms', 'roughly', 'corresponding', 'oseen', 'approximation', 'used', 'asymptotic', 'downstream', 'region', 'second', 'section', 'contains', 'explicit', 'solutions', 'specific', 'configurations', 'particular', 'flows', 'whose', 'initial', 'isovels', 'elliptic', 'shape', 'flows', 'may', 'wakelike', 'jetlike', 'compressibility', 'admitted', 'however', 'flows', 'must', 'uniform', 'pressure', 'must', 'steady', 'final', 'section', 'deals', 'interpretation', 'evaluation', 'results']

```
# Q2) ii) 5
blanks_df = pd.DataFrame(columns = ['Doc Id','Before Blank Space Tokens Removal',
                                    'After Blank Space Tokens Removal'])
removeBlanks(path2,blanks_df,pucntuation_df)
blanks_df.set_index('Doc Id',inplace = True)
blanks_df.head()
```

[23] ✓ 3.5s Python

Doc Id	Before Blank Space Tokens Removal	After Blank Space Tokens Removal
cranfield0690	['investigaion', 'flow', 'spiked-nose', 'hemisphere', 'cylinder', 'mach', 'number', '6', '8.', 'shape', 'nature', 'flow', 'spiked-nose', 'hemispherecylinder', 'studied', 'detail', 'nominal', 'mach', 'number', '6.8', 'reynolds', 'number', 'range', 'based', 'diameter', 'stream', 'conditions', 'ahead', 'model', '0.12', 'x', '10', '1.5', 'x', '10', 'schlieren', 'photographs', 'showed', 'effect', 'varying', 'spike', 'length', 'reynolds', 'number', 'upon', 'shape', 'separated', 'boundary', 'upon', 'location', 'transition', 'heat', 'transfer', 'pressure', 'distribution', 'body', 'correlated', 'location', 'start', 'separation', 'location', 'reattachment', 'location', 'start', 'transition']	['investigaion', 'flow', 'spiked-nose', 'hemisphere', 'cylinder', 'mach', 'number', '6', '8.', 'shape', 'nature', 'flow', 'spiked-nose', 'hemispherecylinder', 'studied', 'detail', 'nominal', 'mach', 'number', '6.8', 'reynolds', 'number', 'range', 'based', 'diameter', 'stream', 'conditions', 'ahead', 'model', '0.12', 'x', '10', '1.5', 'x', '10', 'schlieren', 'photographs', 'showed', 'effect', 'varying', 'spike', 'length', 'reynolds', 'number', 'upon', 'shape', 'separated', 'boundary', 'upon', 'location', 'transition', 'heat', 'transfer', 'pressure', 'distribution', 'body', 'correlated', 'location', 'start', 'separation', 'location', 'reattachment', 'location', 'start', 'transition']
cranfield1184	['three', 'dimensional', 'effects', 'viscous', 'wakes', 'three-dimensionality', 'wakelike', 'jetlike', 'free', 'mixing', 'may', 'stem', 'initial', 'geometric', 'configurations', 'nonuniformities', 'flow', 'variables', 'cross', 'section', 'boundary', 'conditions', 'along', 'flow', 'may', 'generated', 'bodies', 'angle', 'attack', 'nonaxisymmetric', 'bodies', 'mixing', 'nonaxisymmetric', 'jets', 'outer', 'flow', 'finite', 'wings', 'artificial', 'means', 'paper', 'devoted', 'studies', 'bearing', 'configurations', 'first', 'section', 'deals', 'general', 'mathematical', 'model', 'boundary', 'layer', 'approximations', 'used', 'methods', 'solution', 'laminar', 'turbulent', 'flow', 'compressibility', 'unsteadiness', 'streamwise', 'pressure', 'gradients', 'admitted', 'initially', 'flux', 'forms', 'equations', 'given', 'algebraic', 'integrals', 'energy', 'equations', 'diffusion', 'frozenflow', 'equations', 'obtained', 'simplification', 'convective', 'terms', 'roughly', 'corresponding', 'oseen', 'approximation', 'used', 'asymptotic', 'downstream', 'region', 'second', 'section', 'contains', 'explicit', 'solutions', 'specific', 'configurations', 'particular', 'flows', 'whose', 'initial', 'isovels', 'elliptic', 'shape', 'flows', 'may', 'wakelike', 'jetlike', 'compressibility', 'admitted', 'however', 'flows', 'must', 'uniform', 'pressure', 'must', 'steady', 'final', 'section', 'deals', 'interpretation', 'evaluation', 'results']	['three', 'dimensional', 'effects', 'viscous', 'wakes', 'three-dimensionality', 'wakelike', 'jetlike', 'free', 'mixing', 'may', 'stem', 'initial', 'geometric', 'configurations', 'nonuniformities', 'flow', 'variables', 'cross', 'section', 'boundary', 'conditions', 'along', 'flow', 'may', 'generated', 'bodies', 'angle', 'attack', 'nonaxisymmetric', 'bodies', 'mixing', 'nonaxisymmetric', 'jets', 'outer', 'flow', 'finite', 'wings', 'artificial', 'means', 'paper', 'devoted', 'studies', 'bearing', 'configurations', 'first', 'section', 'deals', 'general', 'mathematical', 'model', 'boundary', 'layer', 'approximations', 'used', 'methods', 'solution', 'laminar', 'turbulent', 'flow', 'compressibility', 'unsteadiness', 'streamwise', 'pressure', 'gradients', 'admitted', 'initially', 'flux', 'forms', 'equations', 'given', 'algebraic', 'integrals', 'energy', 'equations', 'diffusion', 'frozenflow', 'equations', 'obtained', 'simplification', 'convective', 'terms', 'roughly', 'corresponding', 'oseen', 'approximation', 'used', 'asymptotic', 'downstream', 'region', 'second', 'section', 'contains', 'explicit', 'solutions', 'specific', 'configurations', 'particular', 'flows', 'whose', 'initial', 'isovels', 'elliptic', 'shape', 'flows', 'may', 'wakelike', 'jetlike', 'compressibility', 'admitted', 'however', 'flows', 'must', 'uniform', 'pressure', 'must', 'steady', 'final', 'section', 'deals', 'interpretation', 'evaluation', 'results']

Question 2:

Methodology -

- First we made a word corpus of all the unique words obtained from the documents after preprocessing.
- Created the inverted index as a python dictionary from the vocabulary obtained above and saved it using pickle.
- For each query input given, first we preprocess the query which includes: a) converting to lowercase b) tokenization c) removing punctuation d) removing special characters e) removing blank spaces.
- Before applying the given operation on query we applied a sanity check to ensure that the number of operations is one less than the number of query terms.
- We took each token and retrieved the postings of each in a list of list.
- Next we scanned the query from left to right and performed any one of the 4 operations: AND, AND NOT, OR and OR NOT.
- Each of the above operations were performed by using 2 pointer approach and a counter was kept for counting each such comparison.
- For more than 1 operation, the resultant of 2 postings was taken to perform the next set of operation with the next token's posting list.
- And lastly the desired output was printed.

Assumption -

While performing OR NOT operation we were supposed to take the compliment of one posting list from the entire set of documents. Thus here we are not considering the number of comparison while computing the compliment set i.e. NOT operation.

Result -

```
n = int(input("Enter number of queries to execute: "))
for i in range(n):
    query = input()
    operation = input()
    retrieve(query,operation,i+1)
```

[28] ✓ 10.2s Python

```
Query 1:  experimental AND flow
Number of documents retrieved for query 1: 158
Names of documents retrieved for query 1: cranfield0001  cranfield0017  cranfield0019  cranfield0025  cranfield0035  cranfield0052  cranfield0053  cranfield0058  cranfield0069
cranfield0070  cranfield0074  cranfield0084  cranfield0103  cranfield0112  cranfield0115  cranfield0121  cranfield0123  cranfield0170  cranfield0171  cranfield0173  cranfield0176
cranfield0179  cranfield0183  cranfield0184  cranfield0186  cranfield0187  cranfield0188  cranfield0189  cranfield0191  cranfield0197  cranfield0206  cranfield0207  cranfield0212
cranfield0216  cranfield0222  cranfield0225  cranfield0227  cranfield0230  cranfield0234  cranfield0257  cranfield0273  cranfield0277  cranfield0283  cranfield0294  cranfield0304
cranfield0307  cranfield0329  cranfield0334  cranfield0344  cranfield0346  cranfield0347  cranfield0360  cranfield0370  cranfield0372  cranfield0377  cranfield0418  cranfield0420
cranfield0421  cranfield0423  cranfield0427  cranfield0435  cranfield0439  cranfield0442  cranfield0453  cranfield0455  cranfield0464  cranfield0467  cranfield0494  cranfield0496
cranfield0498  cranfield0501  cranfield0503  cranfield0504  cranfield0511  cranfield0522  cranfield0536  cranfield0540  cranfield0544  cranfield0549  cranfield0563  cranfield0567
cranfield0569  cranfield0572  cranfield0576  cranfield0588  cranfield0600  cranfield0606  cranfield0610  cranfield0632  cranfield0634  cranfield0635  cranfield0662  cranfield0663
cranfield0666  cranfield0675  cranfield0688  cranfield0689  cranfield0704  cranfield0767  cranfield0772  cranfield0781  cranfield0801  cranfield0802  cranfield0820  cranfield0825
cranfield0856  cranfield0857  cranfield0869  cranfield0907  cranfield0927  cranfield0959  cranfield0965  cranfield0984  cranfield0986  cranfield0996  cranfield1006  cranfield1040
cranfield1074  cranfield1076  cranfield1078  cranfield1080  cranfield1081  cranfield1082  cranfield1083  cranfield1112  cranfield1151  cranfield1153  cranfield1156  cranfield1158
cranfield1159  cranfield1186  cranfield1187  cranfield1195  cranfield1198  cranfield1204  cranfield1205  cranfield1209  cranfield1212  cranfield1213  cranfield1220  cranfield1222
cranfield1225  cranfield1227  cranfield1228  cranfield1230  cranfield1231  cranfield1234  cranfield1262  cranfield1263  cranfield1268  cranfield1269  cranfield1277  cranfield1302
cranfield1310  cranfield1339  cranfield1341  cranfield1374  cranfield1390
Number of comparisons required for query 1: 841
```

```
n = int(input("Enter number of queries to execute: "))
for i in range(n):
    query = input()
    operation = input()
    retrieve(query,operation,i+1)
```

[13] ✓ 12.4s Python

```
Query 1:  jet OR propulsion
Number of documents retrieved for query 1: 92
Names of documents retrieved for query 1: cranfield0007  cranfield0040  cranfield0076  cranfield0077  cranfield0086  cranfield0103
cranfield0126  cranfield0129  cranfield0131  cranfield0137  cranfield0163  cranfield0171  cranfield0172  cranfield0173  cranfield0174
cranfield0176  cranfield0177  cranfield0182  cranfield0218  cranfield0219  cranfield0220  cranfield0243  cranfield0245  cranfield0282
cranfield0290  cranfield0330  cranfield0335  cranfield0341  cranfield0350  cranfield0409  cranfield0453  cranfield0519  cranfield0624
cranfield0636  cranfield0640  cranfield0650  cranfield0692  cranfield0693  cranfield0694  cranfield0695  cranfield0696  cranfield0697
cranfield0721  cranfield0722  cranfield0724  cranfield0725  cranfield0726  cranfield0727  cranfield0729  cranfield0746  cranfield0772
cranfield0773  cranfield0860  cranfield0904  cranfield0908  cranfield0909  cranfield0911  cranfield0946  cranfield0961  cranfield0968
cranfield0969  cranfield0970  cranfield0971  cranfield0972  cranfield0973  cranfield0991  cranfield0992  cranfield0993  cranfield0994
cranfield0997  cranfield1061  cranfield1093  cranfield1098  cranfield1101  cranfield1151  cranfield1195  cranfield1209  cranfield1211
cranfield1212  cranfield1223  cranfield1232  cranfield1244  cranfield1265  cranfield1270  cranfield1292  cranfield1349  cranfield1350
cranfield1351  cranfield1352  cranfield1371  cranfield1374  cranfield1375
Number of comparisons required for query 1: 84
```

Question 3:

Methodology

Bigram inverted index:

Iterate over every possible bigrams in the query using a sliding window of two and then find the list of all possible documents containing that bigrams. While sliding along, go on doing set intersection to find the list of all common docs until that point in the query.

Finally, we end up with list of documents that contains all possible bigrams present in the query.

Positional inverted index:

For two words to be present together in the same doc, both would have their positional indexes with a difference of one ‘in that doc’.

Lets say query = w1 w2 w3…

So, we increment the positional index of w1 for all its docs by 1 and compare it with corresponding docs in w2. If they are one after the other in the same doc, we shall get similar values and intersection would result in a list of those docs. This process is repeated for every new word in the query ahead until we find list of docs which have the exact query within them.

Results:

```
Input[['downstream', 'roughness', 'causes'], ['including', 'gust', 'response']]

Number of documents retrieved for query 1 using bigram inverted index:1
Names of documents retrieved for query 1 using bigram inverted index: ['cranfield0933']
Number of documents retrieved for query 1 using positional inverted index:1
Names of documents retrieved for query 1 using positional inverted index: ['cranfield0933']

Number of documents retrieved for query 2 using bigram inverted index:1
Names of documents retrieved for query 2 using bigram inverted index: ['cranfield0014']
Number of documents retrieved for query 2 using positional inverted index:1
Names of documents retrieved for query 2 using positional inverted index: ['cranfield0014']
```