Importing libraries

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly
remount, call drive.mount("/content/drive", force_remount=True).
```

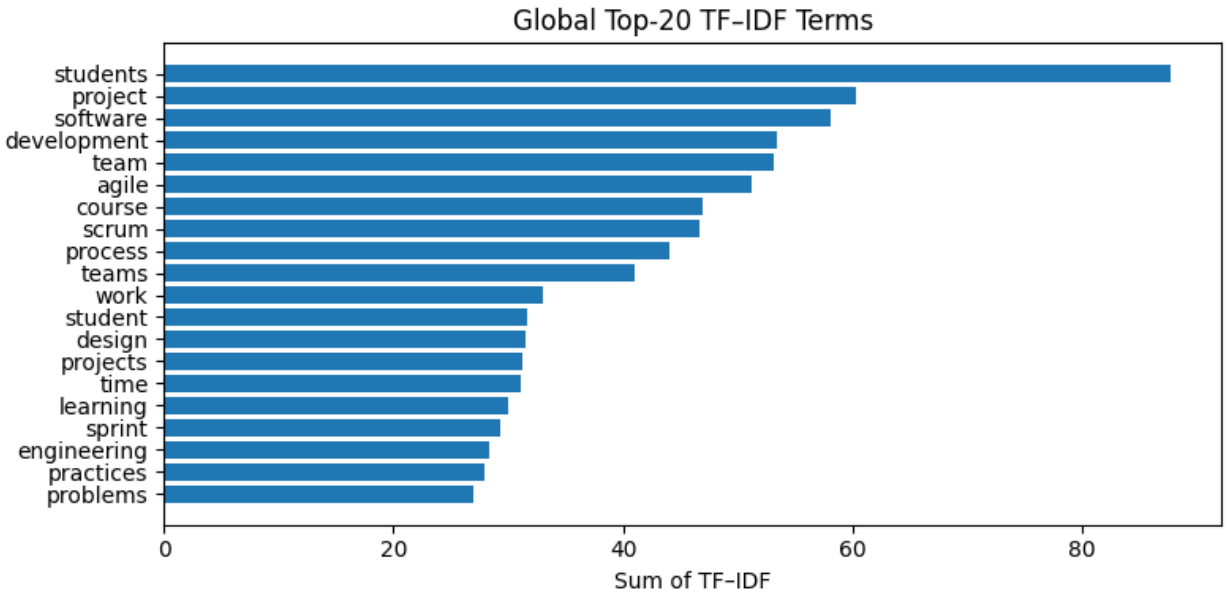
```
import pandas as pd
import re
import matplotlib.pyplot as plt
from wordcloud import WordCloud
from sklearn.feature_extraction.text import TfidfVectorizer,
CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation, PCA
from sklearn.cluster import KMeans
from sklearn.manifold import TSNE
import pyLDAvis
import pyLDAvis.lda_model
```

Data Loading and Preprocessing

```
pos_df = pd.read_csv('/content/drive/MyDrive/VUW
Analysis/positives_compiled.csv')
neg_df = pd.read_csv('/content/drive/MyDrive/VUW
Analysis/negatives_compiled.csv')

# Clean Full Pasage
import re
def clean_text(s):
    s = s.lower()
    s = re.sub(r'\.{2,}', ' ', s)           # remove ellipses
    s = re.sub(r'[^a-z0-9\s]', ' ', s)      # remove punctuation
    s = re.sub(r'\s+', ' ', s).strip()      # collapse whitespace
    s = re.sub(r'[^\x00-\x7F]+', ' ', s)    # only keeps ascii
    return s

pos_df['cleaned'] = pos_df['Full Passage'].apply(clean_text)
neg_df['cleaned'] = neg_df['Full Passage'].apply(clean_text)

# Defining year-group bins (4-year periods, grouping 2016-2019 and
2020-2022)
bins = [2002, 2006, 2010, 2014, 2019, 2022]
labels = ["2003-2006", "2007-2010", "2011-2014", "2016-2019", "2020-
2022"]
pos_df['YearGroup'] = pd.cut(pos_df['Year'], bins, labels=labels,
right=True)
neg_df['YearGroup'] = pd.cut(neg_df['Year'], bins, labels=labels,
right=True)
```

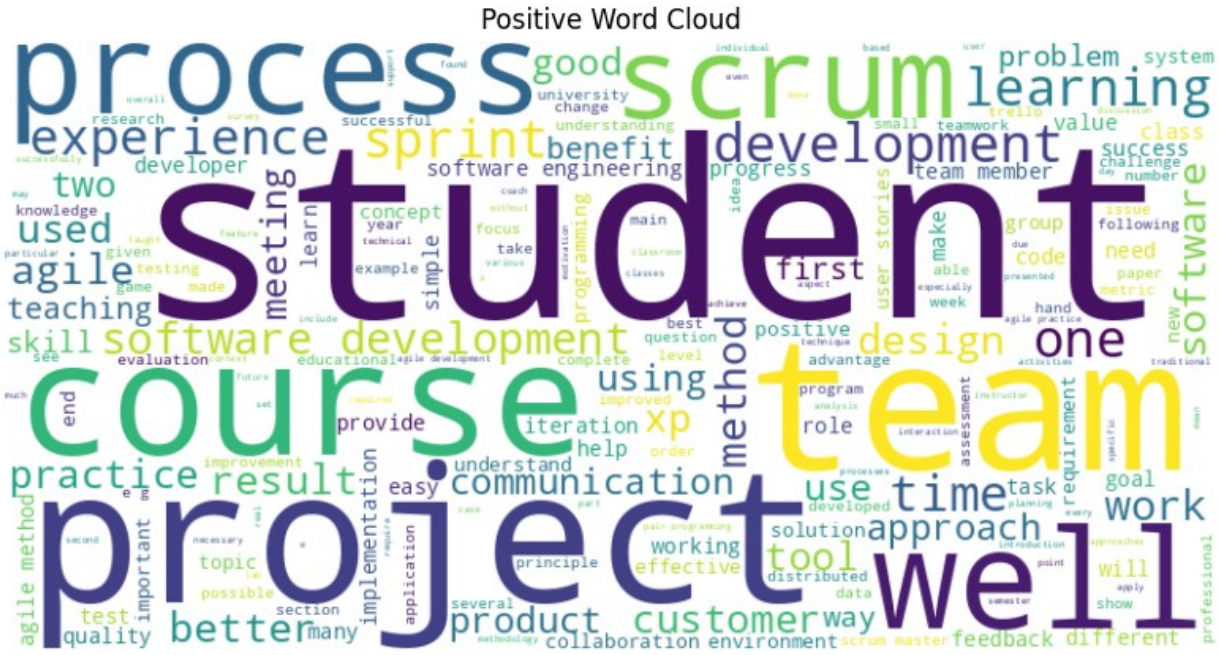
TF–IDF Vectorization

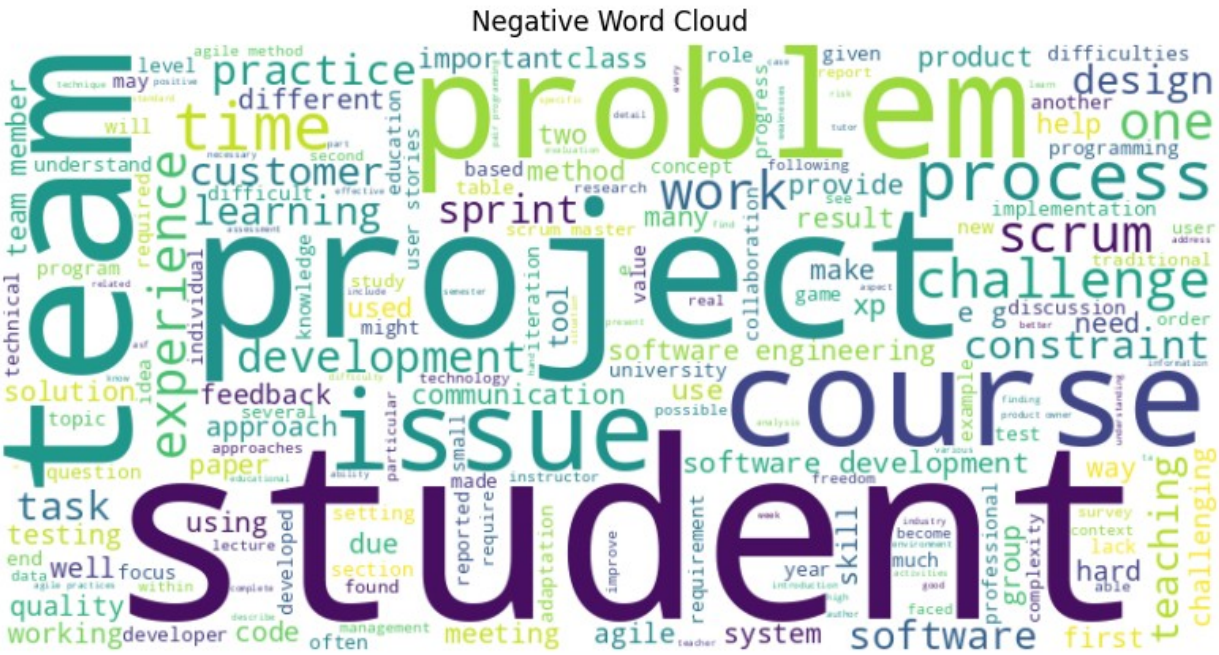
```
# Fitting TF-IDF on combined to get global vocabulary
tfidf = TfidfVectorizer(stop_words='english', max_df=0.9, min_df=2)
tfidf_matrix = tfidf.fit_transform(pd.concat([pos_df['cleaned'],
neg_df['cleaned']]))

# Sum TF-IDF over documents
term_scores = dict(zip(tfidf.get_feature_names_out(),
tfidf_matrix.sum(axis=0).A1))
top_terms = sorted(term_scores.items(), key=lambda x: x[1],
reverse=True)[:20]
terms, scores = zip(*top_terms)

# Bar chart of top terms
plt.figure(figsize=(8,4))
plt.barh(terms[::-1], scores[::-1])
plt.title('Global Top-20 TF–IDF Terms')
plt.xlabel('Sum of TF–IDF')
plt.tight_layout()
plt.savefig('tfidf_top_terms.png')
```




```
for label, df in [('Positive', pos_df), ('Negative', neg_df)]:
    text = ' '.join(df['cleaned'])
    wc = WordCloud(width=800, height=400,
background_color='white').generate(text)
    plt.figure(figsize=(10,5))
    plt.imshow(wc, interpolation='bilinear')
    plt.axis('off')
```

```
    plt.title(f'{label} Word Cloud')
    plt.savefig(f'{label.lower()}_wordcloud.png')
```






```
from sklearn.feature_extraction.text import TfidfVectorizer

# Fit TF-IDF vectorizers
tfidf_pos = TfidfVectorizer(stop_words='english', max_df=0.9,
min_df=2)
```

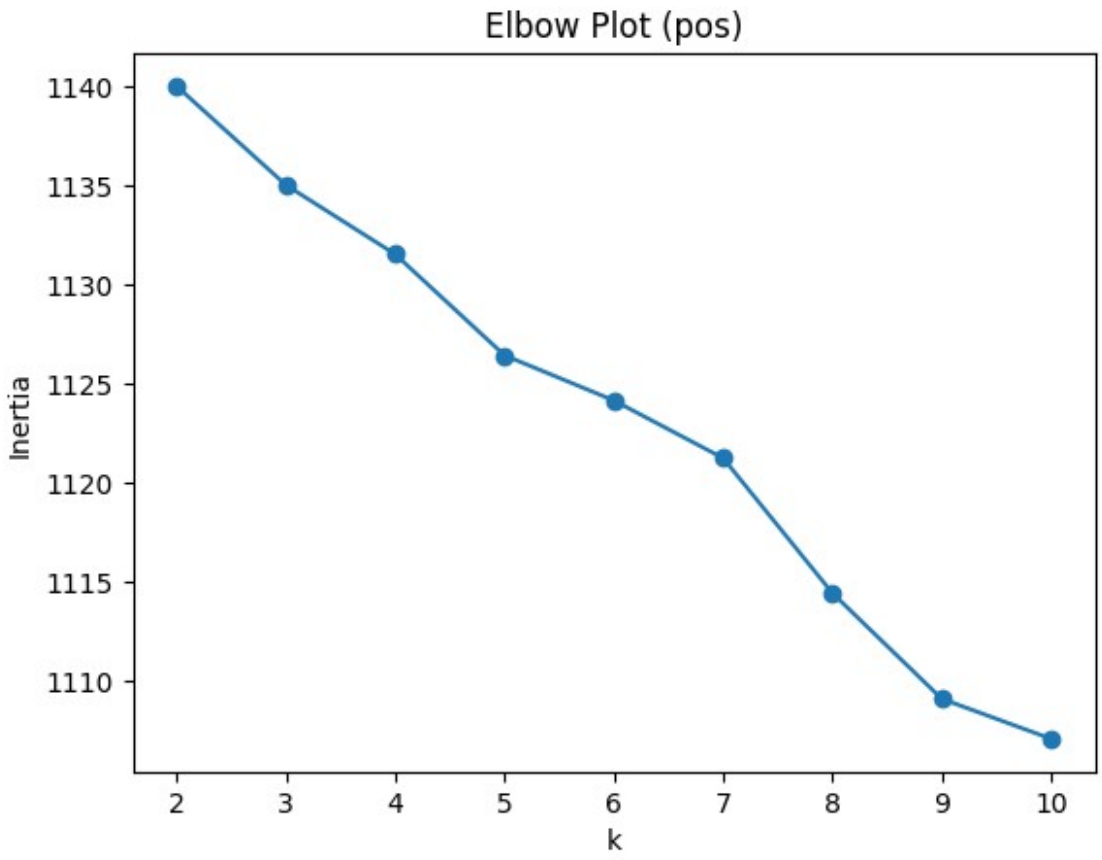
```
tfidf_neg = TfidfVectorizer(stop_words='english', max_df=0.9,
min_df=2)

X_pos = tfidf_pos.fit_transform(pos_df['cleaned'])
X_neg = tfidf_neg.fit_transform(neg_df['cleaned'])
```

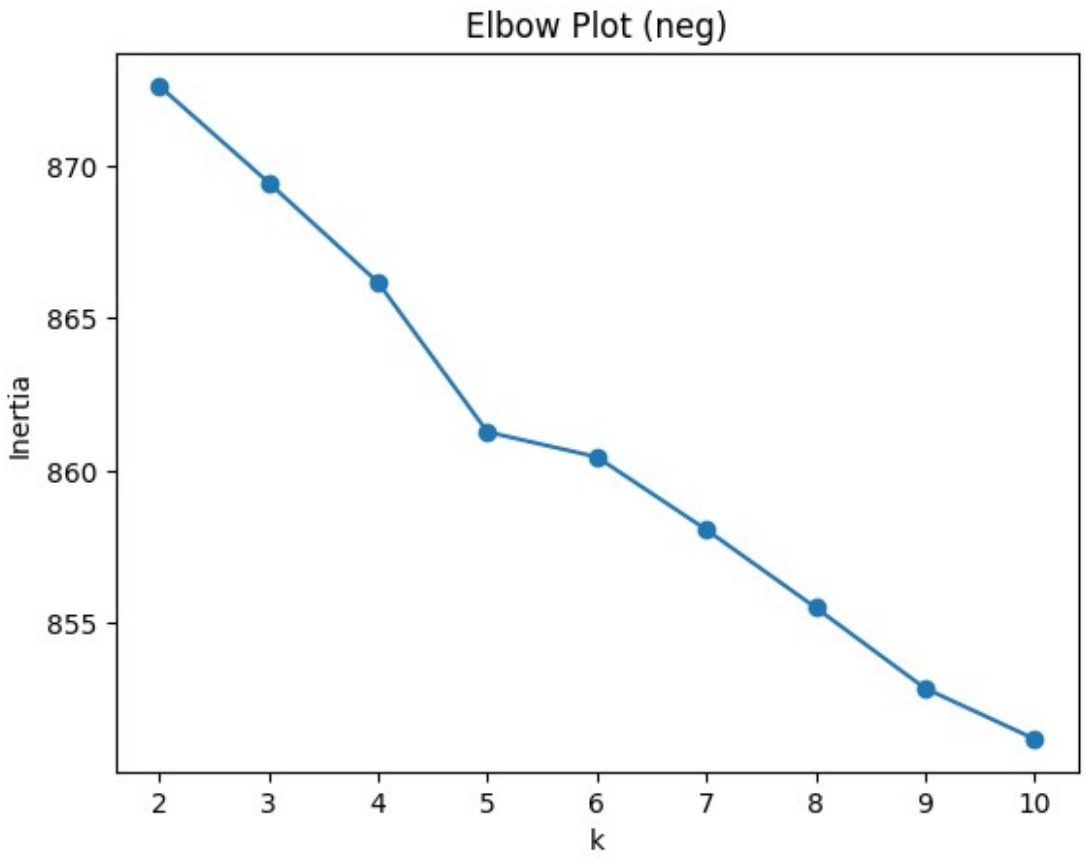
LDA Topic Modeling

```
# Use count vectors for LDA
def fit_lda(df, n_topics=5):
    cv = CountVectorizer(stop_words='english', max_df=0.9, min_df=2)
    Xc = cv.fit_transform(df['cleaned'])
    lda = LatentDirichletAllocation(n_components=n_topics,
random_state=0)
    lda.fit(Xc)
    vis = pyLDAvis.lda_model.prepare(lda, Xc, cv)
    pyLDAvis.save_html(vis, f"lda_{df.name}.html")
    return lda, cv

pos_df.name='pos'
neg_df.name='neg'
lda_pos, cv_pos = fit_lda(pos_df, n_topics=5)
lda_neg, cv_neg = fit_lda(neg_df, n_topics=5)
```

K-Means Clustering

```
# Elbow method for k selection
def plot_elbow(X, max_k=10, prefix='pos'):
    inertias = []
    Ks = range(2, max_k+1)
    for k in Ks:
        km = KMeans(n_clusters=k, random_state=0).fit(X)
        inertias.append(km.inertia_)
    plt.figure()
    plt.plot(Ks, inertias, 'o-')
    plt.xlabel('k')
    plt.ylabel('Inertia')
    plt.title(f'Elbow Plot ({prefix})')
    plt.savefig(f'elbow_{prefix}.png')

plot_elbow(X_pos, max_k=10, prefix='pos')
plot_elbow(X_neg, max_k=10, prefix='neg')

# Fit final clustering (k=5)
km_pos = KMeans(n_clusters=5, random_state=0).fit(X_pos)
pos_df['Cluster'] = km_pos.labels_
km_neg = KMeans(n_clusters=5, random_state=0).fit(X_neg)
neg_df['Cluster'] = km_neg.labels_
```

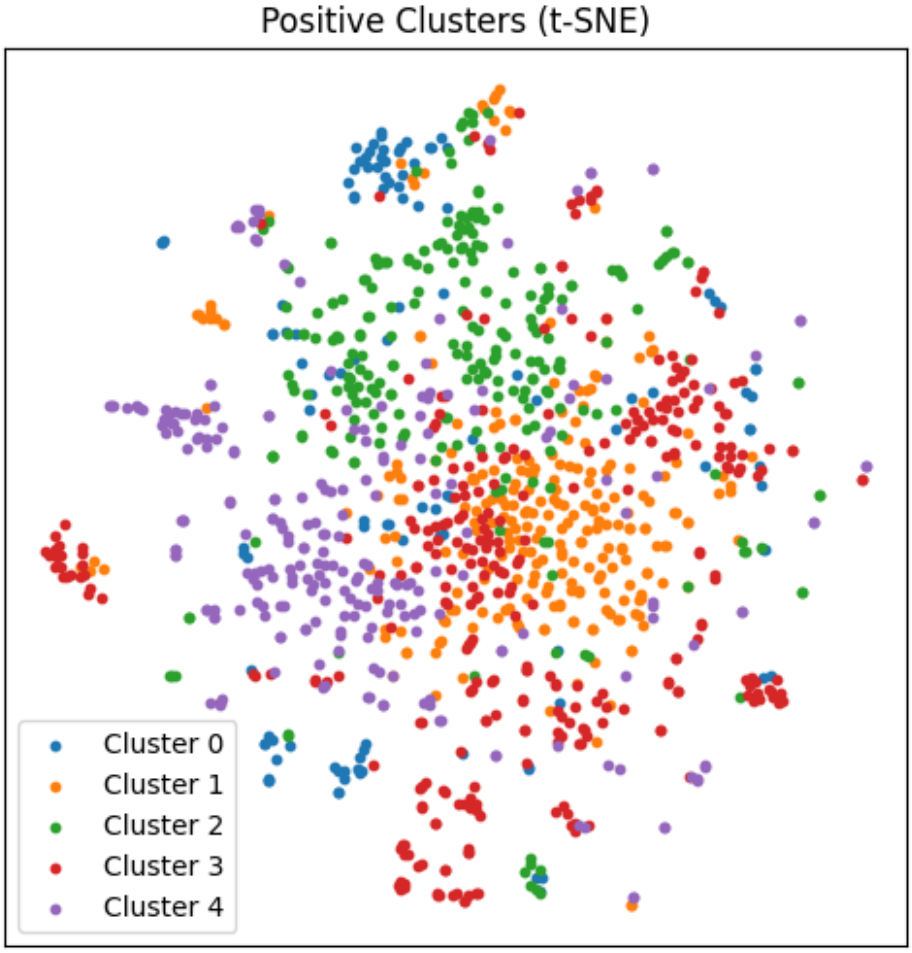

Elbow Plot (pos)
Inertia
1140
1135
1130
1125
1120
1115
1110
2
3
4
5
6
7
8
9
10
k

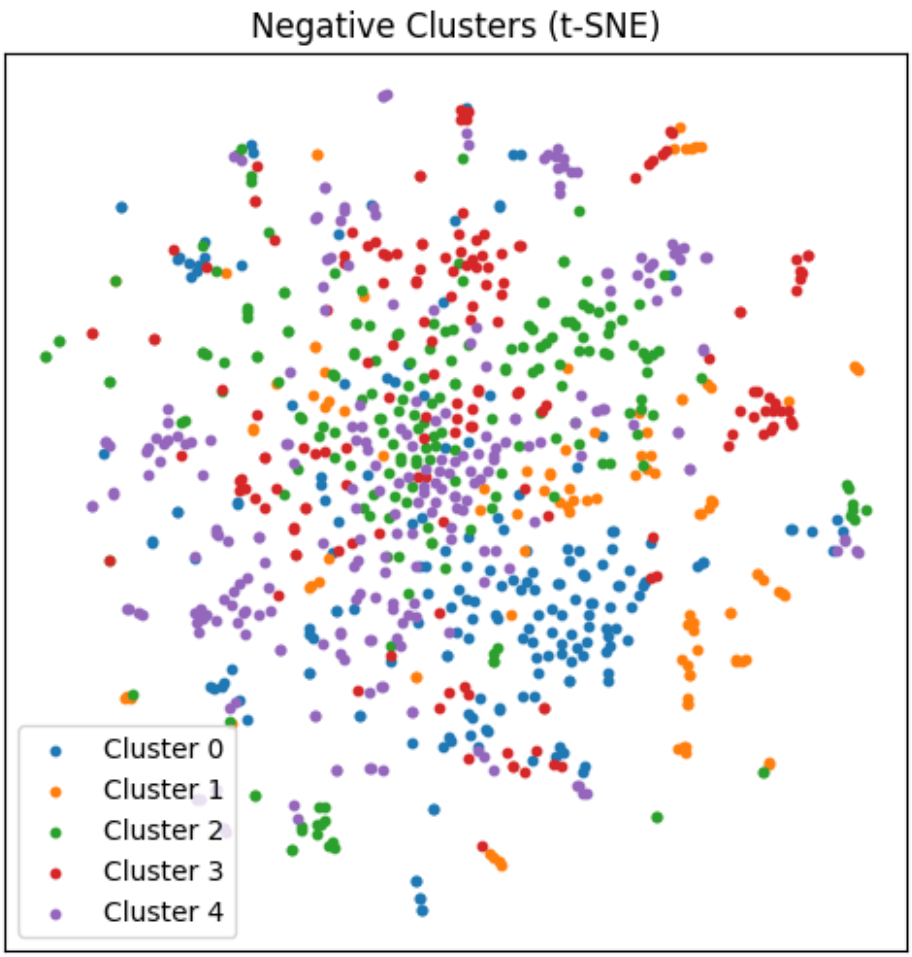


Cluster Visualization (t-SNE & PCA)

```
# Dimensionality reduction + scatter
def plot_embedding(X, labels, title, prefix):
    # PCA to 50D, then t-SNE to 2D for speed
    pca = PCA(n_components=50, random_state=0)
    X_pca = pca.fit_transform(X.toarray())
    tsne = TSNE(n_components=2, random_state=0)
    X_tsne = tsne.fit_transform(X_pca)
    plt.figure(figsize=(6,6))
    for cluster in set(labels):
        idx = labels == cluster
        plt.scatter(X_tsne[idx,0], X_tsne[idx,1], label=f'Cluster
{cluster}', s=10)
    plt.legend()
    plt.title(title)
    plt.xticks([]); plt.yticks([])
    plt.savefig(f'{prefix}_tsne.png')

plot_embedding(X_pos, pos_df['Cluster'].values, 'Positive Clusters (t-
SNE)', 'pos')
plot_embedding(X_neg, neg_df['Cluster'].values, 'Negative Clusters (t-
```

```
SNE)', 'neg')
```


Positive Clusters (t-SNE)
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4



Topic Interpretation & Detailed Category Labeling

```
import numpy as np

def top_topic_words(lda_model, vectorizer, n_top=10):
    feature_names = vectorizer.get_feature_names_out()
    topics = {}
    for i, comp in enumerate(lda_model.components_):
        top_indices = comp.argsort()[-n_top:][::-1]
        topics[i] = [feature_names[idx] for idx in top_indices]
    return topics

# LDA top words
pos_topics = top_topic_words(lda_pos, cv_pos)
neg_topics = top_topic_words(lda_neg, cv_neg)
print("Positive LDA Topics:", pos_topics)
print("Negative LDA Topics:", neg_topics)

# K-Means top TF-IDF terms per cluster
def top_cluster_terms(km, vectorizer, n_top=10):
```

```
    """
    km           : trained KMeans model (e.g. km_pos)
    vectorizer   : the TfidfVectorizer instance used to create X_pos
    n_top        : how many top terms per cluster to return
    """

    feature_names = vectorizer.get_feature_names_out()      # length
must match centroid dims
    centroids     = km.cluster_centers_                     # shape:
(n_clusters, n_features)

    clusters = {}
    for i, centroid in enumerate(centroids):
        # finding the indices of the top-n features for this centroid
        top_indices = centroid.argsort()[-n_top:][::-1]
        # mapping those indices to actual terms
        clusters[i] = [feature_names[idx] for idx in top_indices]
    return clusters



pos_cluster_terms = top_cluster_terms(km_pos, tfidf_pos, n_top=10)
neg_cluster_terms = top_cluster_terms(km_neg, tfidf_neg, n_top=10)

print("Positive Cluster Terms:", pos_cluster_terms)
print("Negative Cluster Terms:", neg_cluster_terms)


# Inspecting a few sample passages for each cluster
for i in range(5):
    print(f"Positive Cluster {i} Samples:")
    print(pos_df[pos_df['Cluster']==i]['cleaned'].sample(3).values)
    print(f"Negative Cluster {i} Samples:")
    print(neg_df[neg_df['Cluster']==i]['cleaned'].sample(3).values)
```

```
Positive LDA Topics: {0: ['students', 'agile', 'design', 'based',
'course', 'development', 'project', 'scrum', 'learning',
'engineering'], 1: ['software', 'development', 'projects', 'agile',
'easy', 'tools', 'real', 'actors', 'r2', 'quality'], 2: ['students',
'project', 'scrum', 'team', 'software', 'agile', 'sprint',
'development', 'course', 'product'], 3: ['process', 'software',
'development', 'students', 'method', 'project', 'team', 'projects',
'programming', 'metrics'], 4: ['students', 'teams', 'student', 'team',
'project', 'communication', 'good', 'work', 'programming', 'scrum']}
Negative LDA Topics: {0: ['team', 'project', 'issues', 'sprint',
'students', 'stories', 'user', 'teams', 'issue', 'scrum'], 1:
['software', 'students', 'agile', 'development', 'process', 'course',
'practices', 'challenges', 'engineering', 'scrum'], 2: ['students',
'problems', 'scrum', 'project', 'constraints', 'team', 'time',
'programming', 'communication', 'meetings'], 3: ['students',
```

```
'software', 'project', 'course', 'design', 'process', 'development',
'problem', 'student', 'time'], 4: ['students', 'teams', 'team',
'projects', 'communication', 'issues', 'challenging', 'customer',
'reported', 'work']}
Positive Cluster Terms: {0: ['customer', 'user', 'time', 'stories',
'story', 'stand', 'usability', 'developers', 'planning', 'teams'], 1:
['students', 'course', 'learning', 'project', 'student', 'real',
'projects', 'feedback', 'skills', 'class'], 2: ['team', 'sprint',
'scrum', 'teams', 'students', 'communication', 'project', 'product',
'members', 'meetings'], 3: ['process', 'design', 'students', 'method',
'code', 'software', 'thinking', 'programming', 'quality',
'development'], 4: ['agile', 'software', 'development', 'project',
'scrum', 'students', 'trello', 'collaboration', 'projects', 'based']}
Negative Cluster Terms: {0: ['software', 'development', 'engineering',
'agile', 'students', 'process', 'project', 'projects', 'programming',
'issues'], 1: ['teams', 'team', 'customer', 'students', 'members',
'project', 'constraints', 'time', 'reported', 'technical'], 2:
['scrum', 'students', 'work', 'team', 'project', 'problems', 'sprint',
'time', 'master', 'meetings'], 3: ['course', 'students', 'agile',
'practices', 'teaching', 'challenges', 'university', 'learning',
'project', 'constraints'], 4: ['problem', 'process', 'design',
'students', 'code', 'team', 'user', 'stories', 'problems', 'project']}
Positive Cluster 0 Samples:
['to try only one idea at a time after a set period they should then
reevaluate their methods to see if what they are trying improved the
product morale and productivity it is important to note that these
comments also show that the students have a definite preference for
extreme programming'
 'point of view of the satisfaction of their two functionalities
quality in terms of unit test coverage readability of code and
experimental stress testing as well as their general usability all
students will be able to distribute points to different teams w r t to
these three measures the number of'
 'tutors pay close attention that user story partitioning is performed
in accordance with recommendations from the literature 22 i e in
features with real costumer value still students tend to create a game
ui kernel user story with the problem that as long as the game ui
kernel is under development']
Negative Cluster 0 Samples:
['to organize a development process how to deal with teams of software
engineers with different skills and motivations and how to produce
outstanding software despite hard deadlines and ideally a 40 hour week
in this paper we report on the setup execution and results of two
software development labs with a'
 'github 24 is a web based git repository which is often used for
hosting open source software projects github provides several
collaboration features such as bug tracking feature requests task
management and wikis for every project 3 trello trello is a very
simple collaboration tool 25 for all kind of projects'
```

'therefore it appears beneficial to share the knowledge acquired ii lab course setup in previous installments of the software development lab course we experienced severe problems often the courses were characterized by a prolonged phase of analysis paralysis i e staying in the analysis phase indefinitely that blended into a hacking']
Positive Cluster 1 Samples:
['prior knowledge the use of appropriate tools that offer increased visibility into the progress of various groups will benefit the instructors both in continual monitoring progress and in feedback provision although there will be some challenges in implementing the proposed structure the benefits of improved student engagement and learning outweigh the'
'laboratory reports 40 final project the main requirement for learning outcomes is their measurability which implies the use of a certain scale for measuring academic achievement in our case in developing the learning outcomes for the theoretical module we used anderson s taxonomy as the one having elaborated vocabulary dictionaries we'
'a project of realistic complexity while working in teams on project tasks students experiment with and reason about the underlying concepts of a discipline potential benefits include active participation in the learning process promotion of critical thinking development of soft skills and a taste of real world projects flipped classroom 11']
Negative Cluster 1 Samples:
['to the customer due to limitations of the webcam when done virtually 5 4 relationship between constraints and adaptations and outcomes applying agile to university projects exhibits constraints such as balancing workload with other courses difficulty in setting up a common time to work together conflicting due dates of deliverables in other'
'to find a time we could coordinate with the product owner to meet for sprint planning retrospectives and reviews 5 1 4 lack of dedication some of the constraints related to dedication of team members for ex ample limited dedication of team members due to other courses or otherwise was reported by half'
'members for example one member of team t10 stated some team members had commitments outside of university such as work or club activities 5 1 6 technical constraintsdifficultyin estimation due to unclear scope was reported by 9 teams 50 t4 t9 t13 t15 t17 similarly two teams specified difference in technical skill level']
Positive Cluster 2 Samples:
['of success in the course students are asked to auto form their team primarily based on project interest interpersonal relationship certainly plays a role as well for students who already know each other 2 project tasks the project tasks are designed based on the dual track agile process presented in section'
'the single team members and the coach possible without the internet and its online collaboration tools distributed student teams would not

be feasible the tools simplify collaboration among students and make progress visible for the coach the same is true for problems and difficulties 1 jenkins ci server jenkins 25 is'
 'student teams to track everyone s contributions and quality of work early in the project students quickly learned to appreciate the team members who did good work and finished what they had promised and took collective pride over the work done i just liked working with the finnish team members']
Negative Cluster 2 Samples:
['final presentation of their developed software project in trello is organised as a game competition to offer them some incentives and encouraging them to work hard as a team and learn from each others work the students are given the opportunity to act in the different roles during the scrum development'
 'three week sprint duration were deemed adequate for project work table 4 feedback on the project work questionnaire item mean std dev project work was challenging 4 255 0 820 i enjoyed working on the project 4 404 0 742 three week duration for each sprint is appropriate 3 787 1'
 'mates with respect to the criteria discussed above supervisors regularly reflect on the progress of the project in retrospective meetings at the end of term problems and solutions are identified reflected and documented in a blog in this way they live the same scrum principles that are taught to students a']
Positive Cluster 3 Samples:
['project but by the end of the project had the same excitement level as the plan driven group lastly while the xp code was structurally better the plan driven group had a much better user interface 6 3 agile topics in our metrics management course our software engineering concentration includes a'
 'f an understanding of professional and ethical responsibility g an ability to communicate effectively h the broad education necessary to understand the impact of engineering solutions in a global economic environmental and societal context i a recognition of the need for and an ability to engage in lifelong learning j a'
 'be prepared for real challenges they may face in developing web applications as many other engineering disciplines 18 19 web application encompasses everything from a simple web page to a comprehensive website 6 web engineering is a framework for building industry quality web applications quick delivery of good quality working web']
Negative Cluster 3 Samples:
['that were created specifically for the evaluation of the new lab course we can deduce that the lab course setup is very well received slight negative feedback was found on imposing the dictum of good enough design i e to only design a system as far as the currently considered user'
 'complete substantial projects until the course is near its end by then there is little time left to complete a project in some

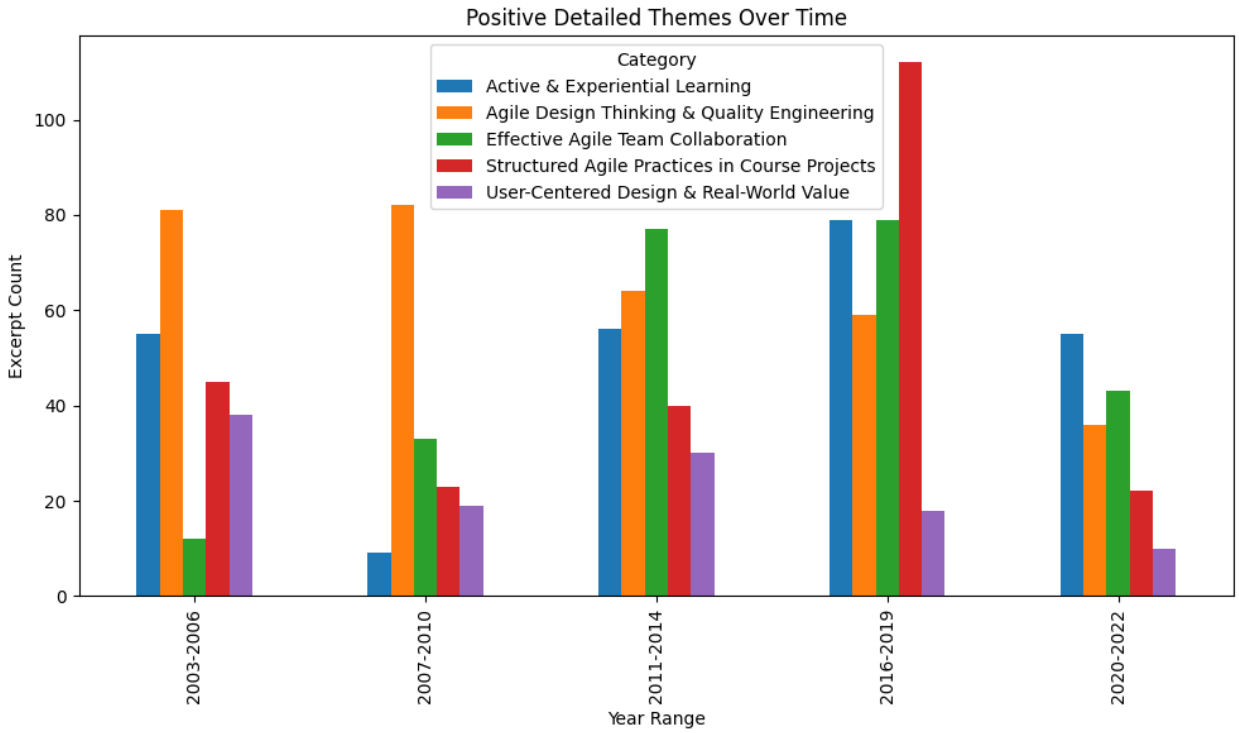
```
curriculums this problem is addressed by a two course sequence a
design is created in the first course and then implemented in the
second regardless of the merits'
 'with little or no time left for implementation consequently students
might be under the impression that the project follows the waterfall
approach in srid this challenge was eventually addressed by framing
the re and ixd practices within dual track agile the introduction of
each task of the innovation project see']
Positive Cluster 4 Samples:
['steps of scrum main roles and artifacts when compared to other
traditional learning processes it is expected from students of this
activity better feelings and good motivations in section ii the scrum
methodology is presented section iii briefly presents some of the
techniques and methods used to teach scrum in'
 'mature agile teams apply more collaboration practices therefore it
is important that the students learn about collaboration practices and
how to live them this is best realized in a real agile project this
paper discusses how educational projects can be setup in such a way
that students can experience the importance'
 'agile experts 24 b use of agile practices in course project an
example the example of a joint utilization of code reviews unit
testing and simple design agile practices in one of the course
projects is described below team based project assignment using at
least 3 different programming languages select from']
Negative Cluster 4 Samples:
['of the backlog the application should include a fully functional
useful usable and delightful front end implemented with html css js
however because of time constraints the back end might include hard
coded functionality and data the righthand side of figure 6 shows a
screen of the mobile application developed by'
 'be implemented scope was adjusted to provide a useful subset of
features and the interface design had to be adjusted to reduce its
scope and complexity it also became apparent that uncertainty existed
regarding some behavioral aspects of the program as well as other
details the original groups were dissolved and'
 'customer and the development team into touch and the retrospective
exposes favorable and unfavorable methods and practices however at the
beginning the estimation process is hard for the students we recommend
that the tutors assist the first estimations by partitioning user
stories into tasks and providing solution approaches but do not']

cluster_to_category_pos = {
    0: "User-Centered Design & Real-World Value",
    1: "Active & Experiential Learning",
    2: "Effective Agile Team Collaboration",
    3: "Agile Design Thinking & Quality Engineering",
    4: "Structured Agile Practices in Course Projects"
}

cluster_to_category_neg = {
```

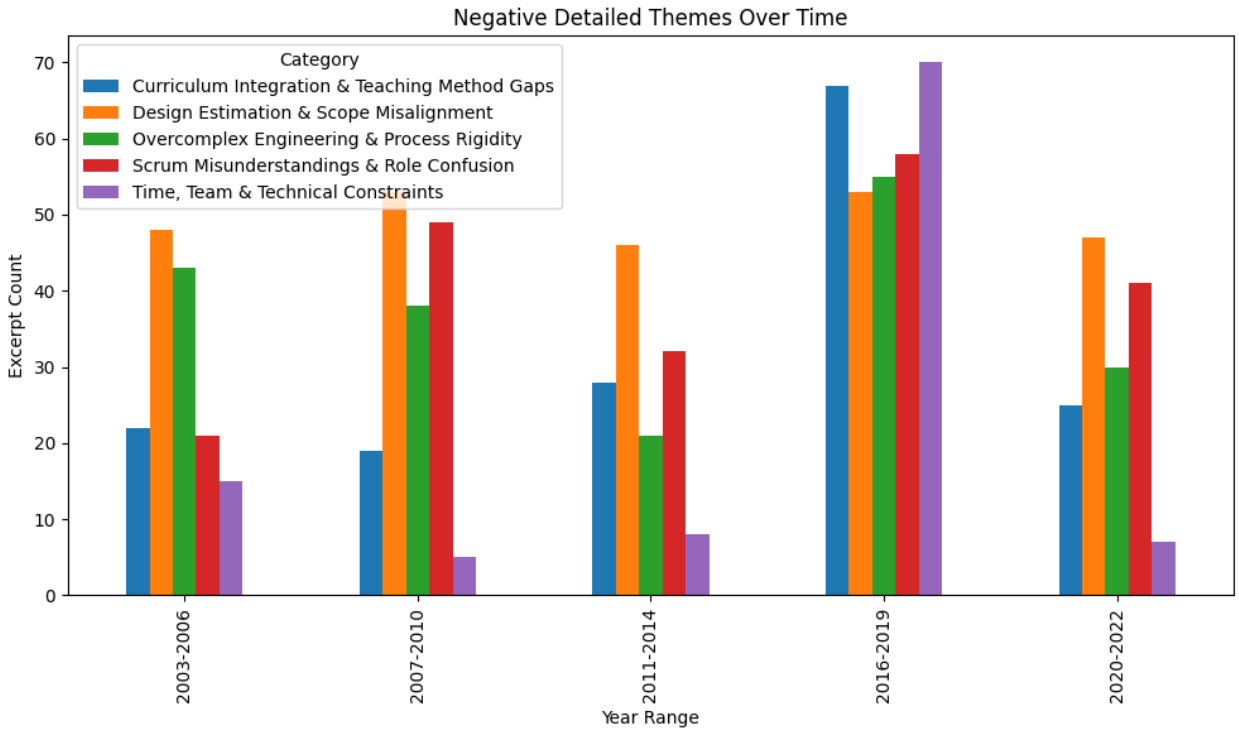
```
    0: "Overcomplex Engineering & Process Rigidity",
    1: "Time, Team & Technical Constraints",
    2: "Scrum Misunderstandings & Role Confusion",
    3: "Curriculum Integration & Teaching Method Gaps",
    4: "Design Estimation & Scope Misalignment"
}

# Apply mappings
pos_df['Category'] = pos_df['Cluster'].map(cluster_to_category_pos)
neg_df['Category'] = neg_df['Cluster'].map(cluster_to_category_neg)
```

Temporal Trend Analysis and Visualization

```
# Count by YearGroup and Category
pos_trends =
pos_df.groupby(['YearGroup','Category']).size().unstack(fill_value=0)
neg_trends =
neg_df.groupby(['YearGroup','Category']).size().unstack(fill_value=0)

# Plot positive detailed trends
pos_trends.plot(kind='bar', figsize=(10,6))
plt.title('Positive Detailed Themes Over Time')
plt.ylabel('Excerpt Count')
plt.xlabel('Year Range')
plt.tight_layout()
plt.savefig('positive_detailed_trends.png')

# Plot negative detailed trends
neg_trends.plot(kind='bar', figsize=(10,6))
plt.title('Negative Detailed Themes Over Time')
plt.ylabel('Excerpt Count')
plt.xlabel('Year Range')
plt.tight_layout()
plt.savefig('negative_detailed_trends.png')
```

```
/tmp/ipython-input-41-3367704716.py:2: FutureWarning: The default of
observed=False is deprecated and will be changed to True in a future
version of pandas. Pass observed=False to retain current behavior or
observed=True to adopt the future default and silence this warning.
  pos_trends =
pos_df.groupby(['YearGroup','Category']).size().unstack(fill_value=0)
/tmp/ipython-input-41-3367704716.py:3: FutureWarning: The default of
observed=False is deprecated and will be changed to True in a future
version of pandas. Pass observed=False to retain current behavior or
observed=True to adopt the future default and silence this warning.
  neg_trends =
neg_df.groupby(['YearGroup','Category']).size().unstack(fill_value=0)
```





Saving results

```
# Saving the fully annotated dataframes
pos_df.to_csv('positive_categorized_detailed.csv', index=False)
neg_df.to_csv('negative_categorized_detailed.csv', index=False)
```