lda_earnings_calls

September 29, 2021

1 Topic Modeling with Earnings Call Transcripts

1.1 Imports & Settings

[2]:
```
%matplotlib inline
import warnings
from collections import Counter
from pathlib import Path

import numpy as np
import pandas as pd
from scipy import sparse

# Visualization
import matplotlib.pyplot as plt
from matplotlib.ticker import FuncFormatter, ScalarFormatter
import seaborn as sns
import ipywidgets as widgets
from ipywidgets import interact, FloatRangeSlider

# spacy for language processing
import spacy
from spacy.lang.en.stop_words import STOP_WORDS

# sklearn for feature extraction
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.feature_extraction import stop_words
from sklearn.model_selection import train_test_split
from sklearn.externals import joblib

# gensim for topic models
from gensim.models import LdaModel
from gensim.models import CoherenceModel
from gensim.corpora import Dictionary
from gensim.matutils import Sparse2Corpus

# topic model viz
import pyLDAvis
```

```
from pyLDAvis.gensim import prepare

# evaluate parameter settings
import statsmodels.api as sm
```

[3]:
```
plt.style.use('fivethirtyeight')
pyLDAvis.enable_notebook()
warnings.filterwarnings('ignore')
pd.options.display.float_format = '{:,.2f}'.format
```

[4]:
```
PROJECT_DIR = Path().cwd().parent.parent
earnings_path = PROJECT_DIR / '03_alternative_data' / '02_earnings_calls' /␣
 ↪'transcripts' / 'parsed'
experiment_path = Path('experiments')
clean_text = Path('data', 'clean_text.txt')
```

[5]:
```
stop_words = set(pd.read_csv('http://ir.dcs.gla.ac.uk/resources/
 ↪linguistic_utils/stop_words',
                             header=None,
                             squeeze=True))
```

1.2 Load Earnings Call Transcripts

The document are the result of scraping the SeekingAlpha Earnings Transcripts as described in n Chapter 3 on Alternative Data.

The transcripts consist of individual statements by company representative, an operator and usually a Q&A session with analysts. We will treat each of these statements as separate documents, ignoring operator statements, to obtain 22,766 items with mean and median word counts of 144 and 64, respectively (or as many as you were able to scrape):

[8]:
```
documents = []
for transcript in earnings_path.iterdir():
    content = pd.read_csv(transcript / 'content.csv')
    documents.extend(content.loc[(content.speaker!='Operator') & (content.
 ↪content.str.len() > 5), 'content'].tolist())
```

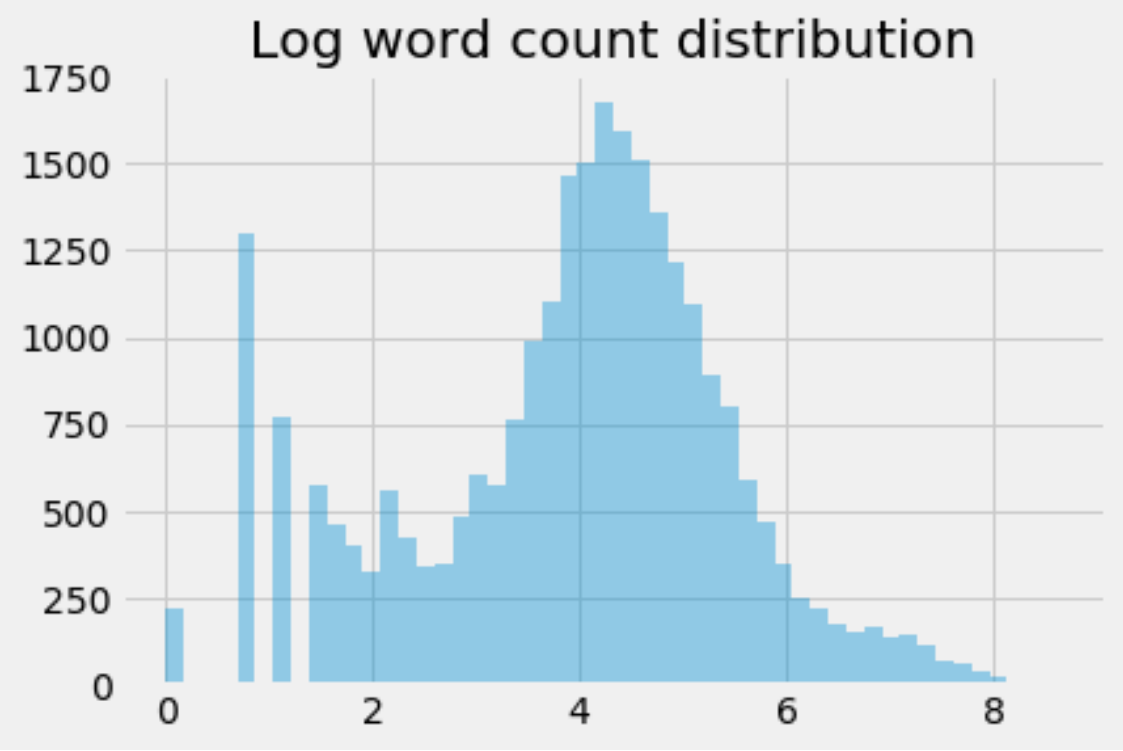
[9]:
```
len(documents)
```

[9]: 26314

1.3 Explore Data

1.3.1 Tokens per document

[10]:
```
word_count = pd.Series(documents).str.split().str.len()
ax = sns.distplot(np.log(word_count), kde=False)
ax.set_title('Log word count distribution');
```



[11]:
```
word_count.describe(percentiles=np.arange(.1, 1.0, .1))
```

[11]:
```
count   26,314.00
mean       139.59
std        287.04
min          1.00
10%          4.00
20%         13.00
30.0%       30.00
40%         46.00
50%         63.00
60%         83.00
70%        113.00
80%        163.00
90%        279.00
max      5,718.00
```

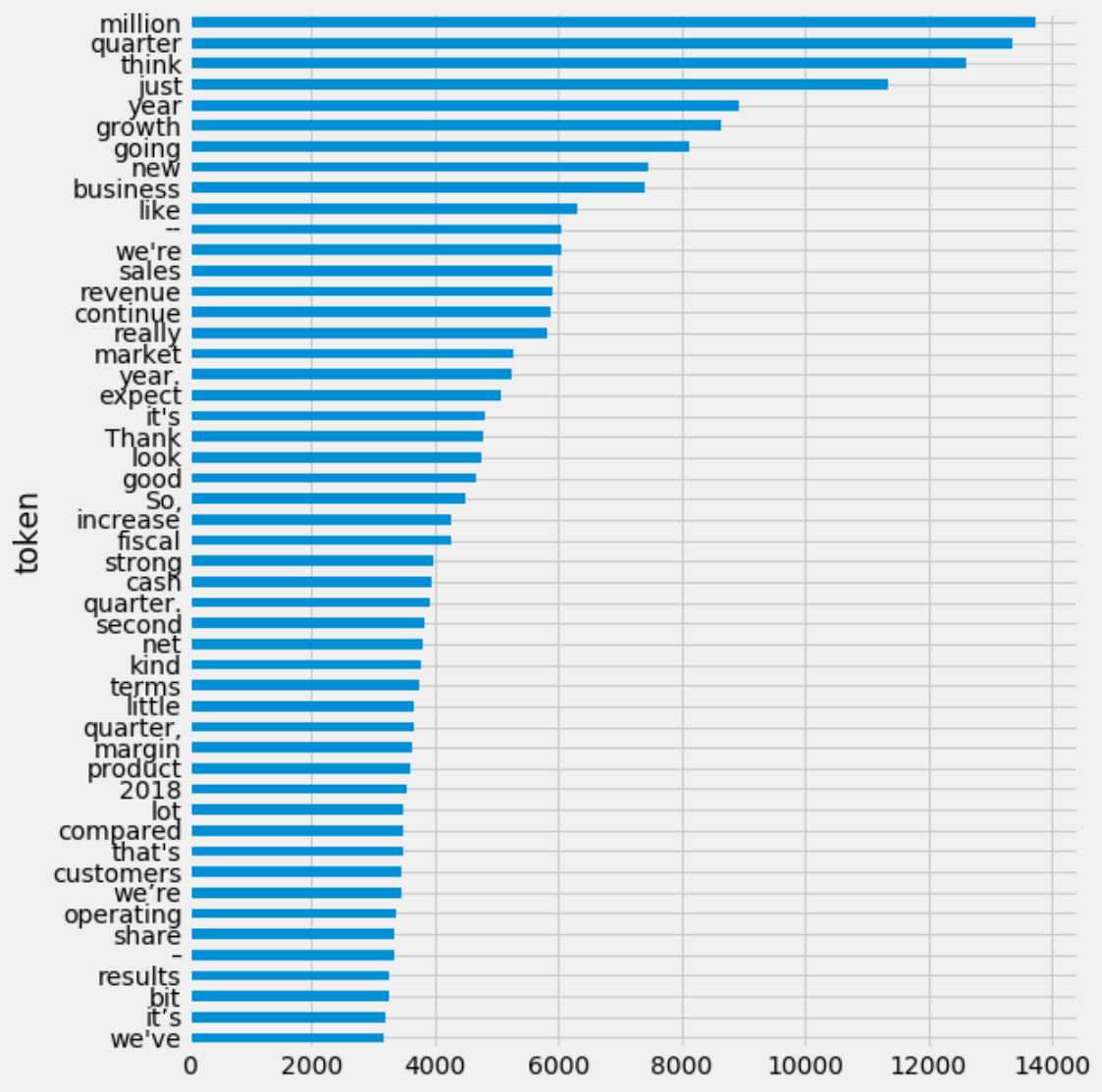
```
dtype: float64
```

[12]:
```
token_count = Counter()
for i, doc in enumerate(documents, 1):
    if i % 5000 == 0:
        print(i, end=' ', flush=True)
    token_count.update(doc.split())
```

```
5000 10000 15000 20000 25000
```

1.3.2 Most frequent tokens

[13]:
```
(pd.DataFrame(token_count.most_common(), columns=['token', 'count'])
 .pipe(lambda x: x[~x.token.str.lower().isin(stop_words)])
 .set_index('token')
 .squeeze()
 .iloc[:50]
 .sort_values()
 .plot
 .barh(figsize=(8, 10)));
```



1.4 Preprocess Transcripts

We use spaCy to preprocess these documents as illustrated in Chapter 13 - Working with Text Data and store the cleaned and lemmatized text as a new text file.

Data exploration reveals domain-specific stopwords like 'year' and 'quarter' that we remove in a second step, where we also filter out statements with fewer than 10 words so that some 16,150 remain.

[14]:
```
def clean_doc(d):
    doc = []
    for t in d:
        if not any([t.is_stop, t.is_digit, not t.is_alpha, t.is_punct, t.
↪is_space, t.lemma_ == '-PRON-']):
```

```
            doc.append(t.lemma_)
    return ' '.join(doc)
```

[17]:
```
nlp = spacy.load('en')
clean_docs = []
for i, document in enumerate(documents, 1):
    if i % 1000 == 0:
        print(f'{i/len(documents):.2%}', end=' ', flush=True)
    doc = nlp(document)
    cleaned = clean_doc(doc)
    if len(cleaned) > 0:
        clean_docs.append(cleaned)
```

```
3.80% 7.60% 11.40% 15.20% 19.00% 22.80% 26.60% 30.40% 34.20% 38.00% 41.80%
45.60% 49.40% 53.20% 57.00% 60.80% 64.60% 68.40% 72.20% 76.01% 79.81% 83.61%
87.41% 91.21% 95.01% 98.81%
```

[18]:
```
clean_text.write_text('\n'.join(clean_docs))
```

[18]:
```
12133756
```

1.5 Vectorize data

[19]:
```
docs = []
for line in clean_text.read_text().split('\n'):
    line = [t for t in line.split() if t not in stop_words]
    if len(line) > 10:
        docs.append(' '.join(line))

len(docs)
```

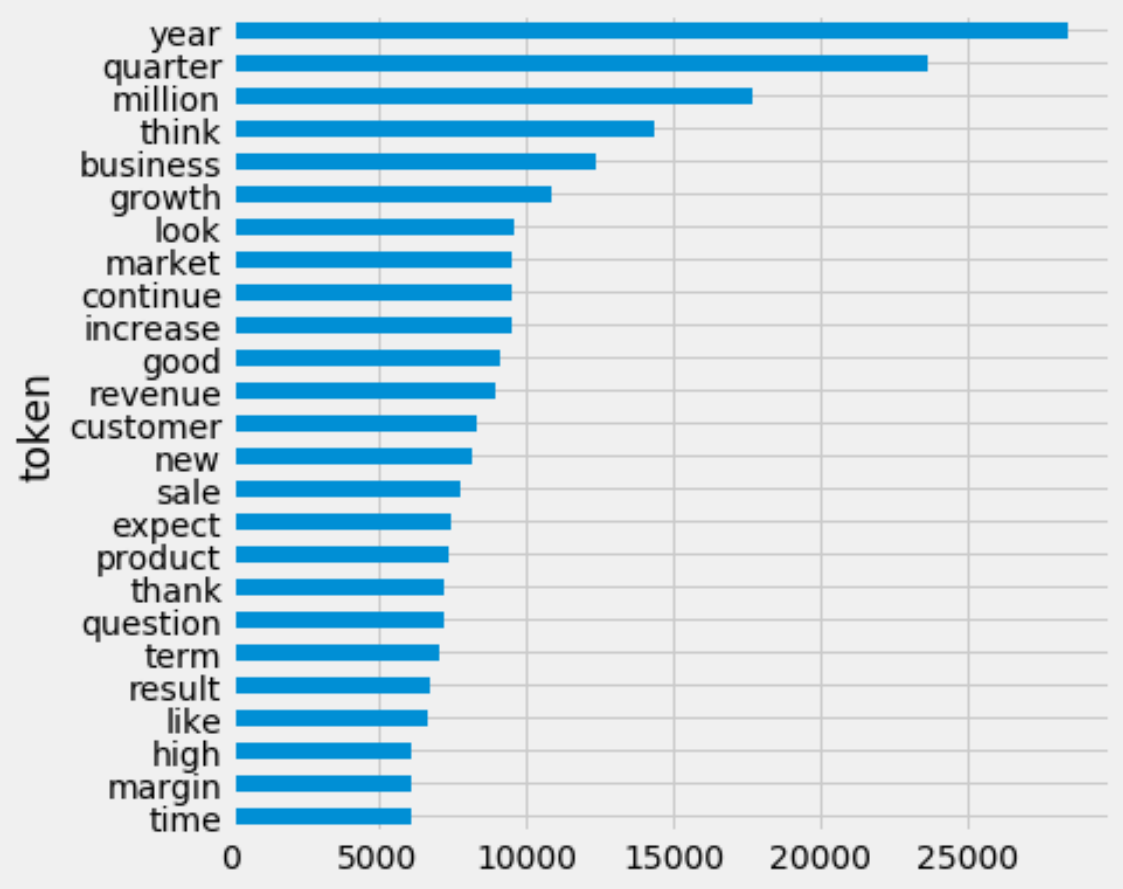
[19]:
```
18600
```

[20]:
```
token_count = Counter()
for i, doc in enumerate(docs, 1):
    if i % 5000 == 0:
        print(i, end=' ', flush=True)
    token_count.update(doc.split())
token_count = pd.DataFrame(token_count.most_common(), columns=['token',␣
↪'count'])
```

```
5000 10000 15000
```

[21]:
```
(token_count
 .set_index('token')
 .squeeze()
 .iloc[:25]
```

```
 .sort_values()
 .plot
 .barh(figsize=(6, 6)));
```



[22]:
```
frequent_words = token_count.head(50).token.tolist()
binary_vectorizer = CountVectorizer(max_df=1.0,
                                    min_df=1,
                                    stop_words=frequent_words,
                                    max_features=None,
                                    binary=True)

binary_dtm = binary_vectorizer.fit_transform(docs)

n_docs, n_tokens = binary_dtm.shape
doc_freq = pd.Series(np.array(binary_dtm.sum(axis=0)).squeeze()).div(binary_dtm.
 ↪shape[0])
max_unique_tokens = np.array(binary_dtm.sum(axis=1)).squeeze().max()
```

[23]:
```
df_range = FloatRangeSlider(value=[0.0, 1.0],
                            min=0,
                            max=1,
                            step=0.0001,
                            description='Doc. Freq.',
                            disabled=False,
                            continuous_update=True,
                            orientation='horizontal',
                            readout=True,
                            readout_format='.1%',
                            layout={'width': '800px'})

@interact(df_range=df_range)
def document_frequency_simulator(df_range):
    min_df, max_df = df_range
    keep = doc_freq.between(left=min_df, right=max_df)
    left = keep.sum()

    fig, axes = plt.subplots(ncols=2, figsize=(14, 6))
    updated_dtm = binary_dtm.tocsc()[:, np.flatnonzero(keep)]
    unique_tokens_per_doc = np.array(updated_dtm.sum(axis=1)).squeeze()
    sns.distplot(unique_tokens_per_doc, ax=axes[0], kde=False, norm_hist=False)
    axes[0].set_title('Unique Tokens per Doc')
    axes[0].set_yscale('log')
    axes[0].set_xlabel('# Unique Tokens')
    axes[0].set_ylabel('# Documents (log scale)')
    axes[0].set_xlim(0, max_unique_tokens)
    axes[0].yaxis.set_major_formatter(ScalarFormatter())

    term_freq = pd.Series(np.array(updated_dtm.sum(axis=0)).squeeze())
    sns.distplot(term_freq, ax=axes[1], kde=False, norm_hist=False)
    axes[1].set_title('Document Frequency')
    axes[1].set_ylabel('# Tokens')
    axes[1].set_xlabel('# Documents')
    axes[1].set_yscale('log')
    axes[1].set_xlim(0, n_docs)
#     axes[1].yaxis.set_major_formatter(ScalarFormatter())

    title = f'Document/Term Frequency Distribution | # Tokens: {left:,d} ({left/
 ↪n_tokens:.2%})'
    fig.suptitle(title, fontsize=14)
    fig.tight_layout()
    fig.subplots_adjust(top=.9)
```

```
interactive(children=(FloatRangeSlider(value=(0.0, 1.0), description='Doc. Freq.
 ↪', layout=Layout(width='800px'…
```

1.6 Train & Evaluate LDA Model

[24]:
```
def show_word_list(model, corpus, top=10, save=False):
    top_topics = model.top_topics(corpus=corpus, coherence='u_mass', topn=20)
    words, probs = [], []
    for top_topic, _ in top_topics:
        words.append([t[1] for t in top_topic[:top]])
        probs.append([t[0] for t in top_topic[:top]])

    fig, ax = plt.subplots(figsize=(model.num_topics*1.2, 5))
    sns.heatmap(pd.DataFrame(probs).T,
                annot=pd.DataFrame(words).T,
                fmt='',
                ax=ax,
                cmap='Blues',
                cbar=False)
    fig.tight_layout()
    if save:
        fig.savefig('earnings_call_wordlist', dpi=300)
```

[25]:
```
def show_coherence(model, corpus, tokens, top=10, cutoff=0.01):
    top_topics = model.top_topics(corpus=corpus, coherence='u_mass', topn=20)
    word_lists = pd.DataFrame(model.get_topics().T, index=tokens)
    order = []
    for w, word_list in word_lists.items():
        target = set(word_list.nlargest(top).index)
        for t, (top_topic, _) in enumerate(top_topics):
            if target == set([t[1] for t in top_topic[:top]]):
                order.append(t)

    fig, axes = plt.subplots(ncols=2, figsize=(15,5))
    title = f'# Words with Probability > {cutoff:.2%}'
    (word_lists.loc[:, order]>cutoff).sum().reset_index(drop=True).plot.
↪bar(title=title, ax=axes[1]);

    umass = model.top_topics(corpus=corpus, coherence='u_mass', topn=20)
    pd.Series([c[1] for c in umass]).plot.bar(title='Topic Coherence',␣
↪ax=axes[0])
    fig.tight_layout();
```

[39]:
```
def show_top_docs(model, corpus, docs):
    doc_topics = model.get_document_topics(corpus)
    df = pd.concat([pd.DataFrame(doc_topic,
                                 columns=['topicid', 'weight']).assign(doc=i)
                    for i, doc_topic in enumerate(doc_topics)])

    for topicid, data in df.groupby('topicid'):
```

```
        print(topicid, docs[int(data.sort_values('weight', ascending=False).
↪iloc[0].doc)])
        print(pd.DataFrame(lda.show_topic(topicid=topicid)))
```

1.6.1 Vocab Settings

For illustration, we create a document-term matrix containing terms appearing in between 0.5% and 50% of documents for around 1,560 features.

[27]:
```
min_df = .005
max_df=.5
ngram_range=(1, 1)
binary = False
```

[28]:
```
vectorizer = CountVectorizer(stop_words=frequent_words,
                             min_df=min_df,
                             max_df=max_df,
                             ngram_range=ngram_range,
                             binary=binary)
```

[29]:
```
dtm = vectorizer.fit_transform(docs)
tokens = vectorizer.get_feature_names()
dtm.shape
```

[29]:
```
(18600, 1541)
```

[30]:
```
corpus = Sparse2Corpus(dtm, documents_columns=False)
id2word = pd.Series(tokens).to_dict()
dictionary = Dictionary.from_corpus(corpus, id2word)
```

1.6.2 Model Settings

[31]:
```
num_topics=15
chunksize=2000
passes=25
update_every=None
alpha='auto'
eta='auto'
decay=0.5
offset=1.0
eval_every=None
iterations=50
gamma_threshold=0.001
minimum_probability=0.01
minimum_phi_value=0.01
per_word_topics=False
```

Training a 15 topic model using 25 passes over the corpus takes a bit over two minutes on a 4-core i7. The top 10 words per topic identify several distinct themes that range from obvious financial information to clinical trials (topic 4) and supply chain issues (12).

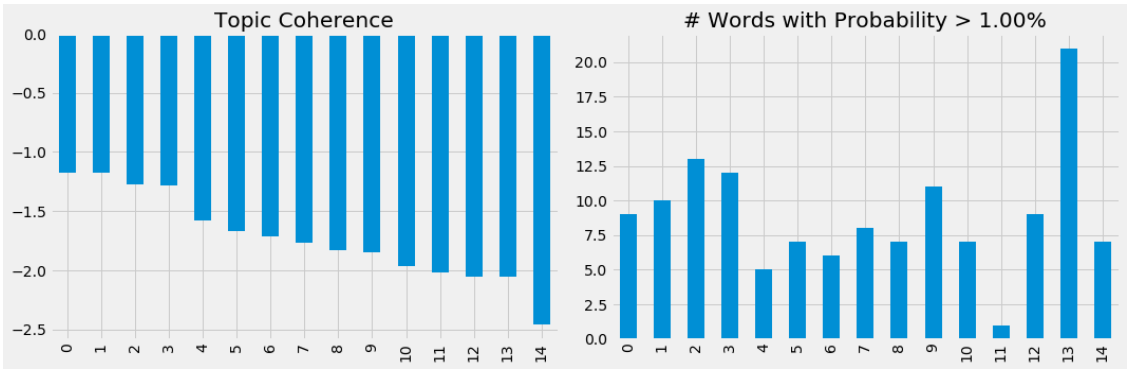
[34]:
```
lda = LdaModel(corpus=corpus,
               id2word=id2word,
               num_topics=num_topics,
               chunksize=chunksize,
               update_every=update_every,
               alpha=alpha,
               eta=eta,
               decay=decay,
               offset=offset,
               eval_every=eval_every,
               passes=passes,
               iterations=iterations,
               gamma_threshold=gamma_threshold,
               minimum_probability=minimum_probability,
               minimum_phi_value=minimum_phi_value,
               random_state=42)
```

[35]:
```
show_word_list(model=lda, corpus=corpus, save=True)
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	expense	statement	basis	focus	cloud	capital	patient	vessel	store	lot	little	project	kind	contract	maybe
1	total	today	tax	deliver	datum	investment	study	fleet	brand	thing	bit	production	sort	fund	guy
2	period	financial	income	service	platform	debt	program	day	category	people	kind	capacity	mean	transaction	okay
3	loss	release	expense	performance	user	asset	clinical	charter	comp	way	price	volume	change	insurance	wonder
4	month	risk	approximately	team	service	flow	trial	ship	inventory	yes	guidance	price	yes	property	kind
5	september	officer	low	strategy	technology	balance	phase	oil	retail	actually	maybe	demand	thing	management	great
6	non	chief	prior	improve	application	loan	datum	fuel	traffic	need	half	low	say	car	guess
7	gaap	gaap	fourth	lead	security	portfolio	development	price	online	different	yes	order	okay	client	follow
8	gross	conference	adjust	provide	enterprise	return	cancer	water	improve	big	say	material	contract	stock	just
9	operating	measure	ebitda	value	solution	sheet	fda	trade	channel	sure	pretty	half	month	vehicle	sort

1.6.3 Topic Coherence

[36]:
```
show_coherence(model=lda, corpus=corpus, tokens=tokens)
```



1.6.4 pyLDAVis

[37]:
```
vis = prepare(lda, corpus, dictionary, mds='tsne')
pyLDAvis.display(vis)
```

[37]: <IPython.core.display.HTML object>

1.6.5 Show documents most represenative of each topic

[40]:
```
show_top_docs(model=lda, corpus=corpus, docs=docs)
```

```
0 excellent thank guy question nice quarter want dig security business assume
real outperformance real highlight recent quarter question think sustainability
growth nice acceleration overall growth decline trend line think sort accelerate
trend line today breadth product sustain accelerate trend line second question
come term investment think key investor debate sort drive high operating margin
able invest business term increase distribution increased product competitive
space grow confidence guy flexibility invest aggressively opportunity thank
        0    1
0   maybe 0.04
1     guy 0.02
2    okay 0.02
3  wonder 0.02
4    kind 0.02
5   great 0.02
6   guess 0.02
7  follow 0.02
8    just 0.01
9    sort 0.01
1 thank rick like mention important point conclude today conference want insight
additional important initiative expect positively impact performance continue
invest key personnel add new marketing sale resource key location globe example
key personnel increase sale activity recent sale win european team support emea
apac region continue growth expand marketing presence attendance important
```

regional event respective sale team continue momentum secondly organization audience trust recommendation gartner analyst mind reengag partnership gartner ensure clear way communication development sale marketing team analyst research team gartner opinion carry weight company look new software vendor happy partner ensure complete understanding technology recommend potential buyer state past maintain close relationship customer result input research vigorously process develop new product new vector revenue broaden market operate specific time competitive reason believe new product revenue vector contribute growth improve customer retention rate contribute increase overall shareholder value inform project progress development year celebrate anniversary company year transformation change make important positive milestone reach prove long term strategy sound reinvent ourself stay true root pioneer original thought leader field service technology market intend continue journey year help customer successful help employee grow succeed career create value shareholder thank attention today continued support like operator open question

	0	1
0	focus	0.01
1	deliver	0.01
2	service	0.01
3	performance	0.01
4	team	0.01
5	strategy	0.01
6	improve	0.01
7	lead	0.01
8	provide	0.01
9	value	0.01

2 thank jack hello positive impact financial follow alignment strategy dedicate resource mass market strategy streamline business deemphasize american academy program lead sequential gross margin expansion fourth quarter quarter sequential reduction million net loss despite million investment large marketing campaign company history support brand effort quarter furthermore expect non tier city expansion strategy support future growth improve financial performance like walk quarter financial highlight net revenue million increase million quarter year increase primarily attribute increase number active student extent increase average revenue active student number active student quarter thousand increase thousand quarter year cost revenue million increase million quarter year increase primarily drive increase total service fee pay teacher mainly delivery increase number pay lesson gross profit million increase million quarter year gross margin compare quarter year total operating expense million increase million quarter year increase mainly result increase sale marketing product development general administrative expense reminder non gaap financial measure exclude share base compensation expense total share base compensation expense million quarter compare million year ago period non gaap sale marketing expense million increase million quarter year increase mainly high branding marketing expense partially offset capitalize sale personnel expense million relation new accounting standard especially asc topic adopt company january clarification new accounting standard thing certain sale commission sale personnel sale agent consider incremental cost obtain contract recognize asset company expect recover

cost adoption new standard million contract cost asset recognize prepaid expense current asset account march include cumulative adjustment million record reduction accumulate deficit reduction sale personnel expense million record sale marketing expense quarter non gaap product development expense million increase million quarter year increase primarily high expense relate technology course development relate personnel strengthen technology platform expand curriculum offering high technical service fee non gaap general administrative expense million increase million quarter year increase primarily result additional expense personnel necessary support expand operation high cost relate compliance reporting obligation public company loss operation million compare million quarter year non gaap loss operation million compare million quarter year foregoing net loss million compare million quarter year non gaap net loss million compare million quarter year basic diluted net loss ads attributable ordinary shareholder compare quarter year ads represent class ordinary share non gaap basic diluted net loss ads attributable ordinary shareholder compare quarter year march company total cash cash equivalent time deposit short term investment million compare million december company current non current deferred revenue billion march compare billion december second quarter currently expect net revenue million million represent increase approximately million quarter year project gross billing million million represent increase approximately million quarter year course outlook base current market condition reflect company preliminary estimate market operating condition customer demand subject change conclude prepared remark open question operator ahead

	0	1
0	expense	0.03
1	total	0.02
2	period	0.02
3	loss	0.02
4	month	0.02
5	september	0.02
6	non	0.02
7	gaap	0.02
8	gross	0.02
9	operating	0.01

3 thank mark good morning slide consolidated sale fourth quarter million prior year stahl add million acquisition sale pass year anniversary stahl acquisition occur january longer acquisition revenue stahl forward exclude fact organic sale growth million sale volume million pricing high previous year basis point overall vertical market remain strong backlog december foreign currency translation continue tailwind increase sale quarter largely result strong euro weak dollar quarter sale million stahl contribute million acquisition revenue sale sale outside million million change fx stahl contribute million acquisition revenue international sale overall solid organic growth quarter emea region strength canada improve emea high single digit organic growth quarter business environment remain strong slide achieve record adjust gross margin quarter year benefit exceptionally strong gross margin stahl quarter likely repeat reflect unusually strong project mix fourth quarter gross profit million increase million adjust gross profit million increase million versus prior year

reconciliation adjust gross profit page presentation let review quarter gross profit bridge stahl acquisition add million adjust gross profit represent stahl gross profit month january high sale volume mix contribute million gross profit positive productivity net cost change plant quarter million increase gross profit foreign currency translation add million gross profit impact high pricing offset raw material inflation positively impact gross profit item positively affect gross profit pro forma item include stahl inventory step expense incur prior year million partial recovery insurance claim add gross profit slide cost million quarter include million pro forma cost relate stahl integration debt repricing fee legal cost insurance recovery litigation exclude item million versus previous guidance million foreign currency translation impact guidance euro strengthen fiscal fourth quarter remainder variance high sell expense fourth quarter partially high sale volume timing certain cost bad debt provision record customer file bankruptcy additional warehouse closure cost relate subleasing compare prior year stahl acquisition add million cost unfavorable foreign currency translation increase cost million incur million high incentive compensation cost cost partially offset million low pro forma item affect year versus year quarterly forecast run rate expect approximate million quarter exclude pro forma item expect incur approximately million additional restructuring action stahl integration yield annual saving approximately million million realize fiscal saving approximately half affect lower spend approximately million quarter begin second quarter estimate total spend stahl integration million saving million fiscal additional million come fiscal turn slide adjust income operation grow million sale compare adjust operating income million prior year adjusted operating margin improve basis point prior year achieve million stahl synergy ahead schedule want point new pension accounting standard effective fiscal impact million pension income annual basis operate income income expense income statement impact net income eps reconciliation adjust operating income page presentation slide gaap earning diluted share versus loss diluted share prior year period adjusted earning diluted share fourth quarter fiscal compare previous year increase share stahl contribute accretion adjust eps quarter add accretion year outstanding result reconciliation gaap earning share adjusted earning share page presentation adjustment tax effect normalized tax rate gaap basis effective tax rate current quarter turn slide year gaap earning diluted share versus diluted share year current year negatively impact effect tax reform adjusted earning diluted share fiscal compare fiscal increase share reconciliation year gaap earning share adjusted earning share page presentation expect year effective tax rate fiscal make significant progress blueprint financial goal adjusted ebitda margin significantly improve prior year return invest capital improve fiscal turn slide work capital percent sale fiscal fourth quarter compare december march working capital percent sale decrease basis point prior year quarter reflect high dpo high accrue liability largely high incentive compensation accrual inventory turn turn low year ago slightly december level carry high inventory level currently improve time delivery market strong backlog substantially believe product line simplification initiative favorably impact turn later fiscal slide net cash operate activity year million high prior year million year free cash flow million guidance capital expenditure fiscal million million turn slide total

debt million net debt million march net debt net total capital repay total million debt year surpass initial target million million set beginning year excellent progress delever achieve net debt adjust ebitda ratio long term target net leverage approximately expect repay million debt fiscal delever balance sheet capital allocation priority continue fund organic growth initiative acquisition consistent blueprint finally return excess cash shareholder dividend share repurchase turn mark wrap

	0	1
0	basis	0.02
1	tax	0.02
2	income	0.02
3	expense	0.02
4	approximately	0.01
5	low	0.01
6	prior	0.01
7	fourth	0.01
8	adjust	0.01
9	ebitda	0.01

4 think supply pretty lock wrong quarter quarter like know frustrate supply forecast past couple year think look broad period time reasonable sense great risk job forecast scenario begin lot history respect gdp job growth typically happen essex market try jump market know geopolitical issue rate rise thing assumption change pretty significantly time change time tend scenario begin strength economy roll mean essex metro sense

	0	1
0	kind	0.03
1	sort	0.03
2	mean	0.02
3	change	0.02
4	yes	0.01
5	thing	0.01
6	say	0.01
7	okay	0.01
8	contract	0.01
9	month	0.01

5 yes flag pick brian work improve execution flagship location traffic location comment persistent traffic headwind location primarily high street tourist location couple thing include roll loyalty program invest store improve conversion benefit able offset headwind traffic long term expect complement location mall base location cultivate particularly brand local customer base drive digital business market digital business remain strong business business wholesale partner region continue prioritize work way flagship build strong base particularly international market

	0	1
0	store	0.06
1	brand	0.03
2	category	0.02
3	comp	0.02

4 inventory 0.01
5   retail 0.01
6   traffic 0.01
7   online 0.01
8   improve 0.01
9   channel 0.01
6 thank martin good nice afternoon think risk point view summary straightforward thing good growth business npl reduce prefer npe ratio npl coverage sum good support economy important low inflow troubled asset good support workout effort let jump page credit risk rwa increase roundabout eur billion growth come come nonretail mainly czech republic romania slovakia good growth execute group corporates markets segment retail increase rwa eur million growth mainly weight bulgaria czech republic slovakia course forget russia martin indicate market risk increase rwe hand hand polish deal transaction forward hedging hedging consume rwa course successful execution transaction fall yes page talk npl provision ratio portfolio level good strong improvement year date talk risk cost npl ratio npl coverage ratio like impairment loss eur million provision ratio basis point think conclude look decent let jump page presentation talk npl distribution country time ask guidance come npl ratio aim easy market way obviously poland albania croatia ukraine confident meet finish year end long question open floor question ahead
            0   1
0   capital 0.03
1 investment 0.02
2   debt 0.02
3   asset 0.02
4   flow 0.01
5   balance 0.01
6   loan 0.01
7   portfolio 0.01
8   return 0.01
9   sheet 0.01
7 good afternoon thank time join today teleconference prepared remark brief past quarter foreseeable future purely execute highly define publicly disclose development plan therapeutic platform asset past month effort simply focus run streamlined efficient clinical corporate operational organization process ph time explore opportunity enhance composite company value proposition announcement appointment oppenheimer september serve strategic advisor capacity oppenheimer work behalf delmar identify evaluate wide range strategic opportunity specific goal facilitate shareholder value generation regard drug development effort fundamental objective continue rapidly efficiently advance phase biomarker drive clinical trial mgmt unmethylat gbm md anderson cancer center houston texas sun yat sen university cancer center guangzhou china addition explore reason financial resource perspective potential treat solid tumor additional oncology indication turn clinical trial respect open label second line avastin naïve study conduct md anderson cancer center pleased report quarter trial continue enroll fast pace originally forecast october trial label mdacc study enrol plan total patient recall rationale second line recurrent gbm

trial initiate february base fact approximately newly diagnose gbm patient grossly underserved current therapy tumor unmethylated mgmt promoter correlate high expression dna repair enzyme mgm scientifically establish patient tumor exhibit high expression mgm t poor prognosis significantly short progression free survival overall survival comparison patient unmethylated mgmt promoter low mgmt expression currently approve therapy goal md anderson study straightforward determine treatment improve overall survival patient progress temozolomide compare historical control base recent increase enrollment rate continue forecast enrollment year end plus month earlier originally plan forecast caveat thanksgiving christmas holiday slightly impact patient enrollment rate phase trial line newly diagnose mgmt unmethylat gbm patient conduct sun yat sen university cancer center guangzhou china commence september similar md anderson study pleased report quarter experience increase rate enrollment october enrol total patient study reminder patient trial treat combination radiotherapy potential alternative current standard care temozolomide radiation regimen currently line treatment patient population operational standpoint trial conduct term collaboration guangxi wuzhou pharmaceutical company study principal clinical goal confirm safety day day dose regimen combination radiotherapy investigate progression free overall survival outcome combination radiotherapy mgmt unmethylat patient date dose confirm cohort study complete base dose conformation phase study select mg metre square combination radiation treatment newly diagnose mgmt unmethylat gbm patient trial addition status phase study want provide brief update publication week sno conference annual meeting society neurooncology meeting abstract present delmar involve phase study provide update md anderson study poster entitle phase study patient mgmt unmethylat bevacizumab naïve recurrent glioblastoma note report october patient enrol patient receive cycle date cut patient currently receive treatment follow survival deceased far study subject receive median cycle therapy patient receive cycle patient receive cycle treatment patient complete cycle therapy subject complete cycle treatment exhibit stable disease end cycle include initially receive starting dose mg metre square initially receive starting dose mg metre square provide update china study poster title phase study radiation therapy newly diagnose mgmt unmethylat glioblastoma report study enrol plan patient dose mg meter squared choose dose treatment remain patient study sno provide new preclinical datum continue support potential combination study potential treatment additional cancer indication gain additional insight unique mechanism action respect combination study preclinical datum strong vivo anti tumor efficacy mgmt unmethylat temozolomide resistant recurrent gbm effect augment combination avastin provide rationale clinical investigation combination avastin treatment gbm second indication diffuse intrinsic pontine glioma dipg brain cancer impact glial cell base brain present preclinical study highlight combination kinase inhibitor promising new therapeutic strategy child dipg ongoing study continue assess vivo activity explore underlying mechanism action combination therapy strategy finally examine efficacy panel gbm cell isolate newly diagnose gbm patient result inhibit neurosphere formation gbm stem cell affect expression protein phosphoprotein central gbm growth salient important finding protein implicate cancer include gbm summary report effective halt growth panel gbm cell respect breadth treatment potential area particular future dennis brown delmar

cofounder chief scientific officer state represent great opportunity treat multiple solid tumor broad range oncology indication new datum ongoing effort support clinical trial complete national cancer institute continue evaluate additional indication provide good treatment option underserved patient strive fiscally responsible fashion delmar work world class partner md anderson cancer center sun yat sen university cancer center china ovarian advisory board enhance outcome underserved oncology patient goal appreciate trust stakeholder place effort way help target population cost effective efficient effectively possible juncture turn scott praill cfo provide summary financial profile quarter end september scott

	0	1
0	patient	0.03
1	study	0.02
2	program	0.01
3	clinical	0.01
4	trial	0.01
5	phase	0.01
6	datum	0.01
7	development	0.01
8	cancer	0.01
9	fda	0.01

8 consolidation real estate industry accelerate crucial real estate developer team financial partner possess depth real estate industry understanding complete merger acquisition consequent funding need enormous recent tightening finance commercial bank real estate company open accept equity investment share profit equity investor hand enormous numerous quality real estate project market difficult project developer obtain financing bright spot advantage private equity fund launch recently good example jupai capture opportunity trend create active manage real estate equity product believe jupai involve active management fund backend power project management bring high management fee cover income forward

	0	1
0	contract	0.02
1	fund	0.01
2	transaction	0.01
3	insurance	0.01
4	property	0.01
5	management	0.01
6	car	0.01
7	client	0.01
8	stock	0.01
9	vehicle	0.01

9 thank matt good afternoon thank join today earning thanksgiving week begin summary result highlight today launch new cloud data services hat provide market update finally tim detailed review financial update outlook result strong quarter execution pure revenue quarter million grow year year gross margin remain high level operating margin exceed upper end guide range strong result indicative forward traction datum centric architecture strategy outline earlier

year customer increasingly voice clear demand hybrid cloud reality today cloud divide evident storage layer prem cloud storage services vary widely make difficult build application run require customer technology choice prem cloud believe way customer able infrastructure choice base good business line base technology live hybrid cloud world application develop deploy seamlessly private public cloud customer increase flexibility insight launch pure cloud data services today announce new set product run natively public cloud deliver unique hybrid cloud storage solution customer flexibility need turn datum value regardless live new product enable customer build hybrid application run seamlessly cloud leverage consistent storage api service benefit cloud native enterprise customer alike announce significant number important new capability pure cloud data services like focus announce beta availability cloud block store base purity software run natively aws cloud block store industrial strength block storage offering enable mission critical enterprise application run cloud capability expect high end storage array cloud block store bring new storage capability like snapshot replication duplication bear cloud web scale app secondly announce availability cloudsnap deliver cloud base datum protection build right flagship flasharray cloudsnap make easy copy snapshot directly public cloud datum protection application migration use case announce beta availability storreduce cloud duplication engine modern backup software design enable simple backup rapid recovery cost effective datum retention public cloud object storage combine flashblade prem solution provide new architecture flash flash cloud enable rapid restore low cost long term cloud retention lastly cloud data services manage cloud datum management solution fact key asset expansion cloud pure manage product cloud day place manage pure product prem extend seamlessly manage pure product cloud enable end end control hybrid cloud mobility protection pure deliver comprehensive cloud datum solution cloud datum infrastructure long provide storage customer build automate private datum cloud new cloud data services enable customer build powerful hybrid cloud solution pure enable comprehensive cloud datum management allow customer manage datum sit excited early response pure cloud data services partner customer analyst look forward work customer better deliver hybrid cloud hat provide quarter market update hat away

	0	1
0	cloud	0.03
1	datum	0.02
2	platform	0.02
3	user	0.01
4	service	0.01
5	technology	0.01
6	application	0.01
7	security	0.01
8	enterprise	0.01
9	solution	0.01

10 okay let start second question come think important teekay supportive transition industry burn clean fuel think look change tanker fleet estimate change lower sulfur fuel oppose scrubber case concern use scrubber obviously transfer sulfur pollution air ocean view viable long term industry operational

constraint fuel quality issue earlier year roughly ship contaminate bad bunker environment heavy fuel oil main fuel industry concern market high sulfur fuel diminish ship water quality fuel come question certainly availability outside major trading bunker hub singapore rotterdam place like reason stance scrubber use high fuel decision install scrubber quality issue concern mention issue maintain additional equipment etcetera use scrubber question term transition physically lot inquiry refiner oil trader term voyage economic voyage currently undertake today oil place traditionally export main refining center look ask question term vessel availability economic different direction think start physically fix cargo inquiry basis point

	0	1
0	vessel	0.02
1	fleet	0.02
2	day	0.02
3	charter	0.01
4	ship	0.01
5	oil	0.01
6	fuel	0.01
7	price	0.01
8	water	0.01
9	trade	0.01

11 thank good afternoon like thank time join sg quarter conference host today paul galvin chief executive officer mahesh shetty company president chief financial officer paul provide business update cover customer partner announcement mahesh discuss financial result press release result cross wire afternoon pm eastern available company website follow management prepared comment open floor question turn management remember certain statement contain release forward look statement meaning private security litigation reform act statement statement historical fact contain presentation include statement future operation financial position business strategy plan objective management future operation forward look statement case forward look statement identify terminology believe estimate continue anticipate intend plan expect predict potential negative term similar expression base forward look statement largely current expectation projection future event financial trend believe affect financial condition result operation business strategy financial need forward look statement subject number risk uncertainty assumption include set forth filing securities exchange commission sec available rely forward look statement prediction future event assure event circumstance reflect forward look statement achieve occur presentation include non gaap financial measure sg block use certain non gaap financial measure assess business operation reference non gaap financial measure consider addition gaap financial measure consider substitute result present accordance gaap finally conference webcast webcast link available investor relations section website time like turn paul galvin paul floor

	0	1
0	statement	0.04
1	today	0.03
2	financial	0.03
3	release	0.02

```
4        risk 0.02
5     officer 0.02
6       chief 0.02
7        gaap 0.02
8  conference 0.01
9     measure 0.01
12 yeah year learning progress fact grow faster channel fact strong base exist
customer tap good relationship think advantage frankly area instance introduce
beam available specifically short period time open clear good channel think
education able clearly message margin perspective look say team fantastic job
try explore new area think time probably talk ron johnson company enjoy grow
exciting costco united states work yield great result great look good stuff best
buy like channel pretty happy point think thing mention past careful distribute
way venture japan instance think good shape distribution wise continue look
audience shop sure place want reason lead costco work ikea think hold lot
promise fiscal confident base plan able grow direct consumer fast pace rest
business continue invest try better year ab testing team step pretty bullish dtc
           0    1
0        lot 0.02
1      thing 0.02
2     people 0.01
3        way 0.01
4        yes 0.01
5   actually 0.01
6       need 0.01
7  different 0.01
8        big 0.01
9       sure 0.01
13 let market argentina particular think healthy domestic demand believe demand
international domestic market devaluation high cost travel abroad rebound
international demand argentina believe weakness month case brazil think healthy
level demand today domestic brazil think reflect quarter number positive action
domestic brazil strong demand compare year internationally somewhat affected
main challenge international relate important increase capacity europe start
industry level slow capacity growth slight decrease look publish second quarter
capacity europe high double digits industry fleet situation stable term demand
international start slight improvement think big concern today imbalance timid
rebound demand compare capacity strong
            0    1
0     project 0.02
1  production 0.02
2    capacity 0.01
3      volume 0.01
4       price 0.01
5      demand 0.01
6         low 0.01
7       order 0.01
8    material 0.01
```

```
9        half 0.01
14 hey garik joe little bit hopefully catch overall framing organic growth right
basically flat overall growth kind weather impact downside kind mid single core
initiative growth think try upper mid single digits range weather ph offset core
initiative look try break core initiative piece little bit mid single roughly
half price half unit volume growth volume growth combination market growth
traditional market overperformance base growth initiative paul say point look
product category point commercial versus residential view pretty similarly point
time term growth rate total mention unit growth rate volume piece roughly range
combination market growth market overperformance
          0    1
0    little 0.03
1       bit 0.03
2      kind 0.02
3     price 0.01
4  guidance 0.01
5     maybe 0.01
6      half 0.01
7       yes 0.01
8       say 0.01
9    pretty 0.01
```

1.7 Review Experiment Results

To illustrate the impact of different parameter settings, we run a few hundred experiments for different DTM constraints and model parameters. More specifically, we let the min_df and max_df parameters range from 50-500 words and 10% to 100% of documents, respectively using alternatively binary and absolute counts. We then train LDA models with 3 to 50 topics, using 1 and 25 passes over the corpus.

The script `run_experiments.py` lets you train many topic models with different hyperparameters to explore how they impact the results. The script `collect_experiments.py` combines the results into a `results.h5` HDF store.

These results are not included in the repository due to their size, but the results are displayed and you can rerun these experiments with earnings call transcripts or other text documents of your choice.

[229]:
```
with pd.HDFStore('results.h5') as store:
    perplexity = store.get('perplexity')
    coherence = store.get('coherence')
```

[230]:
```
perplexity.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 496 entries, 0 to 15
Data columns (total 8 columns):
vocab_size    496 non-null int64
test_vocab    496 non-null int64
```

```
min_df         496 non-null int64
max_df         496 non-null float64
binary         496 non-null bool
num_topics     496 non-null int64
passes         496 non-null int64
perplexity     496 non-null float64
dtypes: bool(1), float64(2), int64(5)
memory usage: 31.5 KB
```

[231]:
```
coherence.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8370 entries, 0 to 713
Data columns (total 7 columns):
topic          8370 non-null int64
passes         8370 non-null object
num_topics     8370 non-null object
coherence      8370 non-null float64
min_df         8370 non-null int64
max_df         8370 non-null float64
binary         8370 non-null bool
dtypes: bool(1), float64(2), int64(2), object(2)
memory usage: 465.9+ KB
```

1.7.1 Parameter Settings: Impact on Perplexity

[221]:
```
X = perplexity[['min_df', 'max_df', 'binary', 'num_topics','passes']]
X = pd.get_dummies(X, columns=X.columns, drop_first=True)
ols = sm.OLS(endog=perplexity.perplexity, exog=sm.add_constant(X))
model = ols.fit(cov_type='HC0')
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             perplexity   R-squared:                       0.772
Model:                            OLS   Adj. R-squared:                  0.765
Method:                 Least Squares   F-statistic:                     75.71
Date:                Mon, 03 Dec 2018   Prob (F-statistic):          5.53e-116
Time:                        15:34:14   Log-Likelihood:                -2407.9
No. Observations:                 496   AIC:                             4848.
Df Residuals:                     480   BIC:                             4915.
Df Model:                          15
Covariance Type:                  HC0
==============================================================================
=
                    coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
```

```
-
const           158.0331      5.804     27.227      0.000     146.657
169.409
min_df_100      -34.7269      4.675     -7.428      0.000     -43.890
-25.564
min_df_250      -77.8456      4.348    -17.903      0.000     -86.368
-69.323
min_df_500     -108.7279      4.475    -24.295      0.000    -117.500
-99.956
max_df_0.25     -20.5220      4.448     -4.614      0.000     -29.240
-11.804
max_df_0.5      -30.2190      4.287     -7.050      0.000     -38.620
-21.818
max_df_1.0      -29.6129      4.442     -6.666      0.000     -38.319
-20.906
binary_True      40.5022      2.764     14.651      0.000      35.084
45.920
num_topics_5      2.4029      3.747      0.641      0.521      -4.941
9.747
num_topics_7      5.7087      3.566      1.601      0.109      -1.280
12.697
num_topics_10    11.3472      3.353      3.385      0.001       4.776
17.918
num_topics_15    20.6403      3.259      6.332      0.000      14.252
27.029
num_topics_20    30.0500      3.500      8.585      0.000      23.190
36.910
num_topics_25    39.5115      4.040      9.780      0.000      31.593
47.430
num_topics_50    89.8369      9.679      9.282      0.000      70.866
108.808
passes_25       -25.4013      2.789     -9.108      0.000     -30.867
-19.935
==============================================================================
Omnibus:                      459.795   Durbin-Watson:                   1.195
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            19757.525
Skew:                           3.888   Prob(JB):                         0.00
Kurtosis:                      32.926   Cond. No.                         12.1
==============================================================================

Warnings:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

1.7.2 Parameter Settings: Impact on Coherence

[232]:
```
X = coherence.drop('coherence', axis=1)
X = pd.get_dummies(X, columns=X.columns, drop_first=True)
ols = sm.OLS(endog=coherence.coherence, exog=sm.add_constant(X))
model = ols.fit(cov_type='HC0')
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              coherence   R-squared:                       0.665
Model:                            OLS   Adj. R-squared:                  0.663
Method:                 Least Squares   F-statistic:                     237.0
Date:                Mon, 03 Dec 2018   Prob (F-statistic):               0.00
Time:                        15:50:17   Log-Likelihood:                -4925.0
No. Observations:                8370   AIC:                             9980.
Df Residuals:                    8305   BIC:                         1.044e+04
Df Model:                          64
Covariance Type:                  HC0
==============================================================================
=
                    coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-
const           -1.5492      0.023    -66.490      0.000      -1.595
-1.504
topic_1         -0.1076      0.020     -5.501      0.000      -0.146
-0.069
topic_2         -0.2316      0.020    -11.553      0.000      -0.271
-0.192
topic_3         -0.3080      0.019    -16.228      0.000      -0.345
-0.271
topic_4         -0.4257      0.019    -21.936      0.000      -0.464
-0.388
topic_5         -0.4553      0.019    -23.826      0.000      -0.493
-0.418
topic_6         -0.5660      0.021    -27.289      0.000      -0.607
-0.525
topic_7         -0.5650      0.020    -28.721      0.000      -0.604
-0.526
topic_8         -0.6366      0.020    -31.807      0.000      -0.676
-0.597
topic_9         -0.7315      0.022    -32.944      0.000      -0.775
-0.688
topic_10        -0.7038      0.020    -34.972      0.000      -0.743
-0.664
topic_11        -0.7618      0.020    -37.648      0.000      -0.801
```

-0.722
topic_12 -0.8278 0.021 -39.529 0.000 -0.869
-0.787
topic_13 -0.9086 0.024 -37.905 0.000 -0.956
-0.862
topic_14 -1.0062 0.028 -36.165 0.000 -1.061
-0.952
topic_15 -0.9400 0.020 -46.426 0.000 -0.980
-0.900
topic_16 -1.0064 0.021 -48.362 0.000 -1.047
-0.966
topic_17 -1.0891 0.024 -45.451 0.000 -1.136
-1.042
topic_18 -1.2143 0.035 -34.359 0.000 -1.284
-1.145
topic_19 -1.3974 0.051 -27.558 0.000 -1.497
-1.298
topic_20 -1.2093 0.032 -38.068 0.000 -1.272
-1.147
topic_21 -1.3008 0.043 -30.040 0.000 -1.386
-1.216
topic_22 -1.3990 0.048 -28.894 0.000 -1.494
-1.304
topic_23 -1.5555 0.075 -20.642 0.000 -1.703
-1.408
topic_24 -1.8207 0.102 -17.928 0.000 -2.020
-1.622
topic_25 -1.2510 0.027 -47.049 0.000 -1.303
-1.199
topic_26 -1.2884 0.027 -47.200 0.000 -1.342
-1.235
topic_27 -1.3080 0.028 -45.896 0.000 -1.364
-1.252
topic_28 -1.3387 0.029 -46.353 0.000 -1.395
-1.282
topic_29 -1.3818 0.033 -41.600 0.000 -1.447
-1.317
topic_30 -1.4258 0.036 -40.118 0.000 -1.495
-1.356
topic_31 -1.4620 0.037 -39.411 0.000 -1.535
-1.389
topic_32 -1.4920 0.038 -38.818 0.000 -1.567
-1.417
topic_33 -1.5331 0.042 -36.156 0.000 -1.616
-1.450
topic_34 -1.5724 0.045 -35.007 0.000 -1.660
-1.484
topic_35 -1.6058 0.046 -34.745 0.000 -1.696

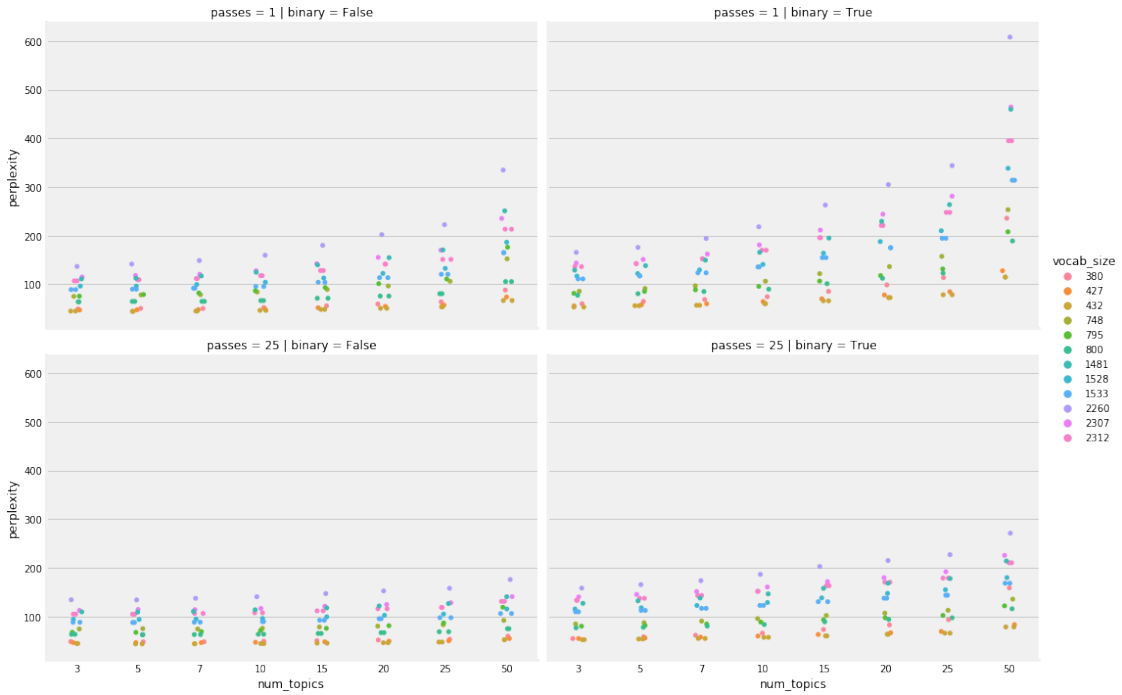
-1.515
topic_36 -1.6599 0.050 -33.476 0.000 -1.757
-1.563
topic_37 -1.7249 0.057 -30.426 0.000 -1.836
-1.614
topic_38 -1.7716 0.061 -28.987 0.000 -1.891
-1.652
topic_39 -1.8252 0.065 -28.099 0.000 -1.953
-1.698
topic_40 -1.8731 0.067 -27.897 0.000 -2.005
-1.741
topic_41 -1.9452 0.073 -26.551 0.000 -2.089
-1.802
topic_42 -2.0258 0.081 -24.863 0.000 -2.186
-1.866
topic_43 -2.1048 0.088 -24.050 0.000 -2.276
-1.933
topic_44 -2.2254 0.103 -21.667 0.000 -2.427
-2.024
topic_45 -2.3408 0.114 -20.463 0.000 -2.565
-2.117
topic_46 -2.4815 0.135 -18.355 0.000 -2.746
-2.217
topic_47 -2.6899 0.161 -16.751 0.000 -3.005
-2.375
topic_48 -3.0070 0.190 -15.830 0.000 -3.379
-2.635
topic_49 -3.3959 0.225 -15.072 0.000 -3.837
-2.954
passes_25 -0.0874 0.010 -9.177 0.000 -0.106
-0.069
num_topics_15 0.0485 0.015 3.204 0.001 0.019
0.078
num_topics_20 0.0827 0.015 5.589 0.000 0.054
0.112
num_topics_25 0.1508 0.015 10.103 0.000 0.122
0.180
num_topics_3 -0.1450 0.029 -4.998 0.000 -0.202
-0.088
num_topics_5 -0.0886 0.021 -4.246 0.000 -0.129
-0.048
num_topics_50 0.4335 0.015 28.089 0.000 0.403
0.464
num_topics_7 -0.0345 0.018 -1.886 0.059 -0.070
0.001
min_df_100 0.2362 0.016 15.220 0.000 0.206
0.267
min_df_250 0.4365 0.016 27.103 0.000 0.405

```
0.468
min_df_500        0.5494      0.016     33.442      0.000       0.517
0.582
max_df_0.25       0.2821      0.014     20.594      0.000       0.255
0.309
max_df_0.5        0.2855      0.013     21.696      0.000       0.260
0.311
max_df_1.0        0.2830      0.014     20.323      0.000       0.256
0.310
binary_True      -0.1248      0.010    -12.994      0.000      -0.144
-0.106
==============================================================================
Omnibus:                     5210.657   Durbin-Watson:                   0.513
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           124064.587
Skew:                          -2.572   Prob(JB):                         0.00
Kurtosis:                      21.146   Cond. No.                         49.5
==============================================================================

Warnings:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

1.7.3 Hyperparameter Impact on Perplexity

[233]:
```
sns.catplot(x='num_topics',
            y='perplexity',
            data=perplexity,
            hue='vocab_size',
            col='binary',
            row='passes',
            kind='strip',
           aspect=1.5);
```

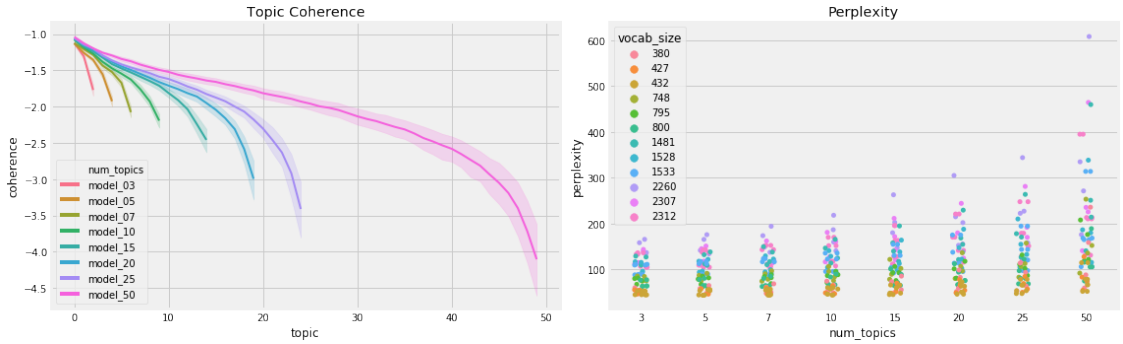


[251]:
```
coherence.num_topics = coherence.num_topics.apply(lambda x: f'model_{int(x):
↪0>2}')
perplexity.min_df = perplexity.min_df.apply(lambda x: f'min_df_{int(x):0>3}')
```

1.7.4 Hyperparameter Impact on Topic Coherence

The following chart illustrate the results in terms of topic coherence (higher is better) ,and perplexity (lower is better). Coherence drops after 25-30 topics, and perplexity similarly increases.

[256]:
```
fig, axes = plt.subplots(ncols=2, figsize=(16,5))
data = coherence.sort_values('num_topics')
sns.lineplot(x='topic', y='coherence', hue='num_topics', data=data, lw=2,␣
↪ax=axes[0])
axes[0].set_title('Topic Coherence')
sns.stripplot(x='num_topics', y='perplexity', hue='vocab_size',␣
↪data=perplexity, lw=2, ax=axes[1])
axes[1].set_title('Perplexity')
fig.tight_layout()
fig.savefig('earnings_call_model_eval', dpi=300);
```



1.8 Load Experiment

The following code let's you load and explore the topic model for a specific experiment run.

1.8.1 Load Document-Term Matrix

[190]:
```
max_df = .1     # [.1, .25, .5, 1.0]
min_df = 250    # [50, 100, 250, 500]
binary= False   # [True, False]
```

[191]:
```
vocab_path = experiment_path / str(min_df) / str(max_df) / str(int(binary))
exp_dtm = sparse.load_npz(vocab_path / f'dtm.npz')
exp_tokens = pd.read_csv(vocab_path / f'tokens.csv', header=None, squeeze=True)
exp_dtm.shape
```

[191]:
```
(22766, 748)
```

[192]:
```
exp_id2word = exp_tokens.to_dict()
exp_corpus = Sparse2Corpus(exp_dtm, documents_columns=False)
exp_dictionary = Dictionary.from_corpus(exp_corpus, exp_id2word)
```

[193]:
```
exp_train_dtm, exp_test_dtm = train_test_split(exp_dtm, test_size=.1)
```

[194]:
```
exp_test_dtm
```

[194]:
```
<2277x748 sparse matrix of type '<class 'numpy.int64'>'
        with 49568 stored elements in Compressed Sparse Row format>
```

[195]:
```
exp_test_corpus = Sparse2Corpus(exp_test_dtm, documents_columns=False)
```

1.8.2 Set Model Parameters

[196]:
```
num_topics = 20 # [3, 5, 7, 10, 15, 20, 25, 50]
passes = 25     # [1, 25]
```

[197]:
```
exp_model_path = vocab_path / str(num_topics) / str(passes)
exp_lda = LdaModel.load(str(exp_model_path / 'lda'))
```

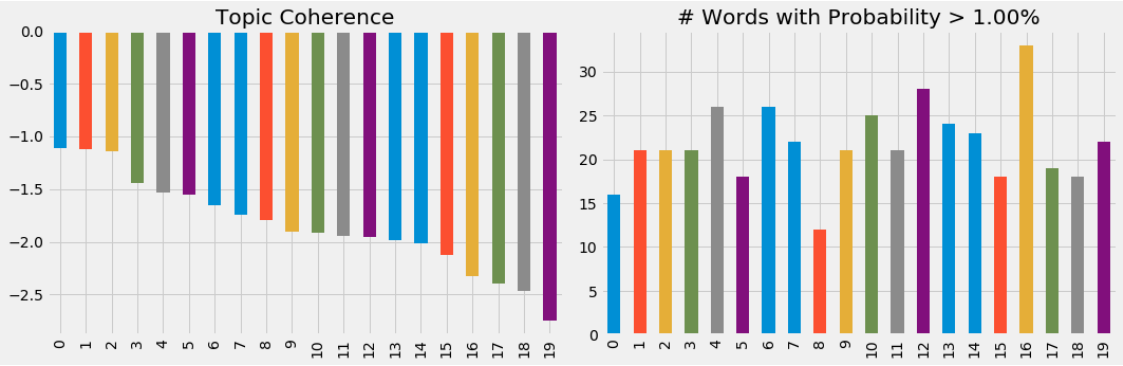
[198]:
```
2 ** (-exp_lda.log_perplexity(exp_test_corpus))
```

[198]: 84.52873752805492

[200]:
```
show_word_list(model=exp_lda, corpus=exp_corpus)
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	statement	compare	strong	rate	cash	brand	patient	service	margin	fiscal	way	price	china	pretty	project	team	store	actually	day	maybe
1	today	expense	in	tax	flow	channel	program	technology	cost	guidance	people	inventory	asset	grow	production	spend	retail	say	today	guy
2	financial	net	focus	basis	capital	consumer	study	cloud	gross	fourth	mean	comp	loan	strong	demand	marketing	open	contract	well	guess
3	release	total	deliver	net	debt	category	development	platform	improvement	ebitda	need	pricing	portfolio	obviously	capacity	content	home	change	appreciate	be
4	risk	period	as	impact	balance	launch	datum	solution	profit	acquisition	opportunity	volume	credit	overall	order	user	location	course	week	sort
5	officer	cash	this	decline	share	online	phase	datum	line	range	try	impact	bank	big	fleet	management	north	impact	month	follow
6	chief	non	performance	earning	sheet	commerce	trial	large	improve	organic	sure	supply	investment	probably	cost	member	season	that	interest	just
7	gaap	loss	grow	segment	return	marketing	test	system	mix	half	different	positive	risk	feel	fuel	help	plan	indiscernible	operator	wonder
8	include	approximately	drive	income	shareholder	digital	complete	partner	drive	estimate	able	weather	client	rate	slide	all	traffic	no	remark	hi
9	information	decrease	believe	low	dividend	experience	process	industry	impact	adjust	obviously	week	management	half	facility	job	america	mean	prepared	take

[201]:
```
show_coherence(model=exp_lda, corpus=exp_corpus, tokens=exp_tokens)
```



[202]:
```
exp_vis = prepare(exp_lda, exp_corpus, exp_dictionary, mds='tsne')
pyLDAvis.display(exp_vis)
```

[202]: <IPython.core.display.HTML object>

1.8.3 Review Documents by Topic

[209]:
```
exp_docs = []
for line in clean_text.read_text().split('\n'):
        exp_docs.append(line)
len(exp_docs)
```

[209]: 22745

[207]:
```
doc_topics = exp_lda.get_document_topics(exp_corpus)
df = pd.concat([pd.DataFrame(doc_topic, columns=['topicid', 'weight']).
 ↪assign(doc=i) for i, doc_topic in enumerate(doc_topics)])
```

[210]:
```
for topicid, data in df.groupby('topicid'):
    print(topicid, exp_docs[int(data.sort_values('weight', ascending=False).
 ↪iloc[0].doc)])
    print(pd.DataFrame(exp_lda.show_topic(topicid=topicid)))
```

```
0 no real update give time continue different clinical program hifu application
so track update give far
            0    1
0     project 0.05
1  production 0.03
2      demand 0.03
3    capacity 0.02
4       order 0.02
5       fleet 0.02
6        cost 0.02
7        fuel 0.02
8       slide 0.02
9    facility 0.02
1 okay understand and second question multipart hopefully quick the ballast
water treatment mention coast guard approval think important can feeling timing
so schedule far number ship year what hire time associate ship fitting ballast
water treatment system cost ship and cost expense amortize
            0    1
0     service 0.04
1  technology 0.04
2       cloud 0.03
3    platform 0.03
4    solution 0.03
5       datum 0.02
6       large 0.02
7      system 0.02
8     partner 0.02
9    industry 0.01
2 thank
             0    1
```

```
0    statement 0.05
1        today 0.03
2    financial 0.03
3      release 0.02
4         risk 0.02
5      officer 0.02
6        chief 0.02
7         gaap 0.02
8      include 0.02
9  information 0.02
3 thank nick as mention outset prepared remark want spend minute look ahead rest
fiscal year aware run rate basis substantially ahead guidance give april quarter
performance contain time revenue contribution increase quarterly revenue state
quarter experience revenue attrition approximately million equate begin revenue
run rate approximately million year for project slow pace attrition equate
revenue base approximately million starting point base timing revenue attrition
occur project quarter low revenue quarter fiscal year begin improvement
quarterly revenue go forward base contribution expect realize evaluator include
sale book month continue forecast total annual revenue million million fiscal
year adjust ebitda contribution approximately million million that conclude
prepared remark turn operator want thank streamline health associate continue
hard work dedication client shareholder turn operator session operator
               0    1
0       actually 0.05
1            say 0.03
2       contract 0.03
3         change 0.03
4         course 0.02
5         impact 0.02
6           that 0.02
7  indiscernible 0.02
8             no 0.02
9           mean 0.01
4 think traffic drive initiative business customer initiative pilot seven store
msa central california early quarter roll store think help drive traffic
business customer specifically household traffic tell main focus initiative
nielsen analytic pricing promotion in addition think make good traction customer
service talk shopping experience store think good wait time good stock think
industry lead think quarter stock actually smart final banner strong stock think
build ticket think build traffic rely item exactly customer look so think
combination thing and course cycle promotional activity start quarter help
stabilize let begin grow traffic
           0    1
0      price 0.11
1  inventory 0.06
2       comp 0.05
3    pricing 0.04
4     volume 0.04
```

```
5     impact 0.04
6     supply 0.03
7   positive 0.02
8    weather 0.02
9       week 0.02
5 so project talk eau du soleil toronto the total award million ship ship fiscal
large project project architecturally specify building product so necessarily
look uptick total sale help solidify expect fiscal business
         0    1
0     rate 0.06
1      tax 0.05
2    basis 0.04
3      net 0.02
4   impact 0.02
5  decline 0.02
6  earning 0.02
7  segment 0.02
8   income 0.02
9      low 0.02
6 okay thank yunlong impressive result quarter so question briefly asset zmn
global education so possible share month quarter quarter trajectory asset
understand quarter second quarter asset actually improve pretty especially line
growth sale expense so possible share bit color asset perform quarter outlook
year
             0    1
0       strong 0.02
1           in 0.02
2        focus 0.02
3      deliver 0.01
4           as 0.01
5         this 0.01
6  performance 0.01
7         grow 0.01
8        drive 0.01
9      believe 0.01
7 and continue able expand add new product delta activewear season get good
response digital print customer so product line continue expand capability able
increase delta product go digital print business
            0    1
0       brand 0.10
1     channel 0.04
2    consumer 0.04
3    category 0.03
4      launch 0.03
5      online 0.03
6    commerce 0.02
7   marketing 0.02
8     digital 0.02
```

9 experience 0.02
8 think early year expect able start talk tumor type work continue work additional pathway breast cancer pathway particular tumor type identify new subtype lead potential drug program instance program hop place c met combination significantly expand range alternative think base current research opportunity potentially increase number pathway study breast cancer proceed number front identify new patient population exist tissue type new tissue type
             0   1
0   patient 0.04
1   program 0.03
2   study 0.02
3 development 0.02
4   datum 0.02
5   phase 0.02
6   trial 0.02
7   test 0.01
8   complete 0.01
9   process 0.01
9 so term revenue growth quarter quarter launch car txpress platform so think contribute new account perspective growth quarter historical revenue grow start expand new account jeff comment
               0   1
0   compare 0.06
1   expense 0.05
2   net 0.04
3   total 0.03
4   period 0.03
5   cash 0.02
6   non 0.02
7   loss 0.02
8 approximately 0.02
9   decrease 0.02
10 and follow comment broker channel say see increase emphasize past wonder look forward expect rely broker channel remain year
            0   1
0   china 0.05
1   asset 0.05
2   loan 0.04
3   portfolio 0.04
4   credit 0.03
5   bank 0.03
6 investment 0.03
7   risk 0.02
8   client 0.02
9 management 0.02
11 yes think right way look right way look
            0   1
0   team 0.09

```
1       spend 0.04
2   marketing 0.04
3     content 0.04
4        user 0.03
5  management 0.03
6      member 0.03
7        help 0.03
8         all 0.03
9         job 0.03
```

12 thank operator want welcome earning conference today in addition wang jingbo chairlady ceo noah cfo shang participate for today agenda briefly summarize noah overall performance quarter development core wealth management asset management business chairlady wang provide current view overall market regulatory environment product strategy cfo shang follow detailed discussion noah quarter financial performance conclude question answer session as enter second half market sentiment remain volatile in trade dispute stringent financial regulatory reform de leveraging investment appetite substantially reduce despite clear policy signal include tax cut encouragement private enterprise development continue open door policy market confidence fragile during transitional period chinese company undergo brutal ph test market this true china wealth asset management industry as lead firm noah continue focus provide high quality service create value client improve risk management core competency maintain sustainable growth prepare challenge face in quarter company business grow steadily perform board in quarter non gaap net income attributable shareholder year year million year year million quarter the group net revenue quarter billion year year in wealth management segment raise billion financial product quarter transaction value quarter year year pleased achieve financial performance challenging environment in wealth management business demand professional wealth management service china strong recruit suitable talent foster steady development particularly important period as end quarter frontline cover city country number relationship manager increase year year in order grow client enhance ability provide comprehensive service noah continue host depth investor education event since beginning year hold hundred event noah company visit small investment session depth client meeting for ultra high net worth client provide customize service activity closed session senior management tailor trust planning service private event black card client through detailed analysis client profile relationship black card client truly drive company enhancement service pure financial product sale for exist client focus heavily improvement post investment service periodic fund performance report comprehensively systemize templat concise easy client read retrieve online fund update information communicate client voice app relationship manager add enhanced human touch integrate client service resource new client operation center in addition exist client service hotline add ai support customer service assistance actively reach exist client better understand need during past month client operation center host op conference gopher externally manage fund believe initiative help relationship manager provide professional post investment service client plan organized manner furthermore research work revamp noah research team lot systematic professional in quarter research team provide

training session frontline professional cover fund manager different asset class publish internal external research report conduct interview domestic international medium a substantial number medium online platform cite research result brand reputation market influence noah research firmly establish active user noah research online channel increase month asset management business grow healthily as end quarter gopher total asset management reach billion increase year year break asset class aum private equity investment increase year year reach billion account total asset management the aum credit real estate secondary market equity discretionary management respectively account total asset management with demand china high net worth client specialized institutionalized gopher continue enhance investment capability breadth depth the demand institutional investor rise development chinese asset management industry gopher recently win bid manager billion fund invest government indiscernible important recognition fund management experience as lead alternative asset manager industry gopher establish standardized process online operation cover entire process fundraising investment management redemption gopher comprehensively sort integrate investment management system through visualize display interface system display paramedic view historic investment database include dozen fund fund hundred investing fund thousand portfolio company strive fully explore value datum resource these initiative integrate resource capital technology operation office support provide strong foundation gopher growth for overseas business high net worth client active global asset allocation noah globalization strategy continue forward steadily office hong kong us canada australia singapore provide diversified product service offering client as end quarter gopher overseas asset management reach billion increase yearly basis as continue expand global business attract overseas investor interested china at end october china australia family office alliance create in alliance noah join hand australia family office open investment opportunity family office country bring cooperation opportunity experienced sharing wealth management estate planning lastly like share group progress technology operation recently intelligent customer service robot launch official website public account noah online app question issue major category include group introduction product service daily conversation resolve ai as end september intelligent robot answer nearly customer question stop noah account system continuously improve audio visual recording client order information contract signing product status update launch online product information deliver customer timely accurately believe embrace technology help noah client service capability year change put high demand company operation management risk control but time noah face harsh market condition believe short term market fluctuation cause external sentiment actually rise opportunity long term return continue adhere strategy build core competency maintain compliance stable management focus growth uncertain market environment with turn noah chairlady ceo wang jingbo speak chinese remark follow english translation

	0	1
0	store	0.25
1	retail	0.04
2	open	0.04
3	home	0.04

```
4  location 0.03
5     north 0.03
6    season 0.03
7      plan 0.03
8   traffic 0.02
9   america 0.02
13 tantan officially roll membership subscription business january number
subscriber grow impressive pace faster pace see momo similar company roll type
service believe speak strong demand come tantan user improve change match
purchasing value add service and think tantan stage early stage term
monetization and look roadmap similar company actually achieve acceleration
number subscriber growth monetization growth up point tantan membership
subscription area think tantan lot potential grow number subscriber monetization
in addition tantan continue grow user base reach social feature think tantan lot
opportunity monetize membership subscription business area diversify business
line obviously momo lot expertise resource tantan leverage road high level
confidence future growth trend tantan pay user operator
             0    1
0          way 0.03
1       people 0.02
2         mean 0.02
3         need 0.02
4  opportunity 0.02
5          try 0.02
6         sure 0.02
7    different 0.02
8         able 0.01
9    obviously 0.01
14 yes previous guidance million billion synergy take billion billion synergy
and consistent expect million synergy year
             0    1
0       fiscal 0.17
1     guidance 0.12
2       fourth 0.08
3       ebitda 0.05
4  acquisition 0.05
5        range 0.04
6      organic 0.03
7         half 0.02
8     estimate 0.02
9       adjust 0.01
15 no see consistent growth similar quarter term aftermarket demand meet
salesperson review customer retirement plan and tell aware increase retirement
fluctuation fuel price certain retirement build model continue operate but
advise change customer utilization result fuel price so continue strengthen
building aftermarket
           0    1
0     pretty 0.03
```

```
1       grow 0.03
2     strong 0.02
3  obviously 0.02
4    overall 0.02
5        big 0.02
6   probably 0.02
7       feel 0.02
8       rate 0.02
9       half 0.01
16 great thank
        0    1
0   maybe 0.06
1     guy 0.04
2   guess 0.04
3      be 0.04
4    sort 0.03
5  follow 0.03
6    just 0.03
7  wonder 0.02
8      hi 0.02
9    take 0.02
17 yes close expect leverage de lever indicate spectrum have equity ownership
outstanding share indicate issue equity future so financing place look give
market condition add additional equity future
             0    1
0         cash 0.09
1         flow 0.06
2      capital 0.06
3         debt 0.04
4      balance 0.04
5        share 0.03
6        sheet 0.02
7       return 0.02
8  shareholder 0.02
9     dividend 0.02
18 well network perform couple record month quarter release so great compare
transaction revenue subscription revenue change dynamic bit but think pretty
frankly but network look couple month highest think largely attributable
customer network good place transact business thing come pretty good number
             0    1
0       margin 0.14
1         cost 0.12
2        gross 0.05
3  improvement 0.03
4       profit 0.02
5         line 0.02
6      improve 0.02
7          mix 0.02
```

```
8        drive 0.02
9       impact 0.02
19 thank join fiscal year earning with today ceo george kurian cfo ron pasek
this webcast live available replay website as reminder adopt new accounting
standard asc historical financial result restate conform new accounting revenue
recognition rule reconciliation previously report gaap result restate gaap
result gaap non gaap result include earning release applicable period post
website financial table guidance historical supplemental datum table non gaap
gaap reconciliation unless note refer non gaap number during today forward look
statement projection respect financial outlook future prospect guidance quarter
fiscal year expectation future revenue profitability cash flow shareholder
return ability grow expand opportunity involve risk uncertainty disclaim
obligation update forward look statement projection actual result differ
materially statement projection variety reason include global political
macroeconomic market condition ability expand total available market introduce
deliver new differentiate product service disruption manage gross profit margin
capitalize market position cloud strategy maintain execution continue capital
allocation strategy please refer document file time time sec available website
specifically recent form fiscal year current report form during financial
measure present non gaap indicate turn george
            0    1
0         day 0.10
1       today 0.04
2        well 0.04
3  appreciate 0.03
4        week 0.03
5       month 0.02
6    interest 0.02
7    operator 0.02
8      remark 0.02
9    prepared 0.02
```

[ ]: