06_lda_earnings_calls

September 29, 2021

1 Topic Modeling with Earnings Call Transcripts

1.1 Imports & Settings

[1]:
```
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
%matplotlib inline

from collections import Counter
from pathlib import Path

import numpy as np
import pandas as pd

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

# spacy for language processing
import spacy

# sklearn for feature extraction
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

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

# topic model viz
import pyLDAvis
from pyLDAvis.gensim_models import prepare

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

```
/home/stefan/.pyenv/versions/miniconda3-latest/envs/ml4t/lib/python3.7/site-
packages/patsy/constraint.py:13: DeprecationWarning: Using or importing the ABCs
from 'collections' instead of from 'collections.abc' is deprecated since Python
3.3,and in 3.9 it will stop working
  from collections import Mapping
```

[3]:
```
sns.set_style('white')
```

[4]:
```
pyLDAvis.enable_notebook()
```

[6]:
```
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

[5]:
```
PROJECT_DIR = Path().cwd().parent
data_path = PROJECT_DIR / 'data' / 'earnings_calls'
```

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):

[7]:
```
documents = []
for i, transcript in enumerate(data_path.iterdir()):
    content = pd.read_csv(transcript / 'content.csv')
    documents.extend(content.loc[(content.speaker!='Operator') & (content.
 ↪content.str.len() > 5), 'content'].tolist())
```

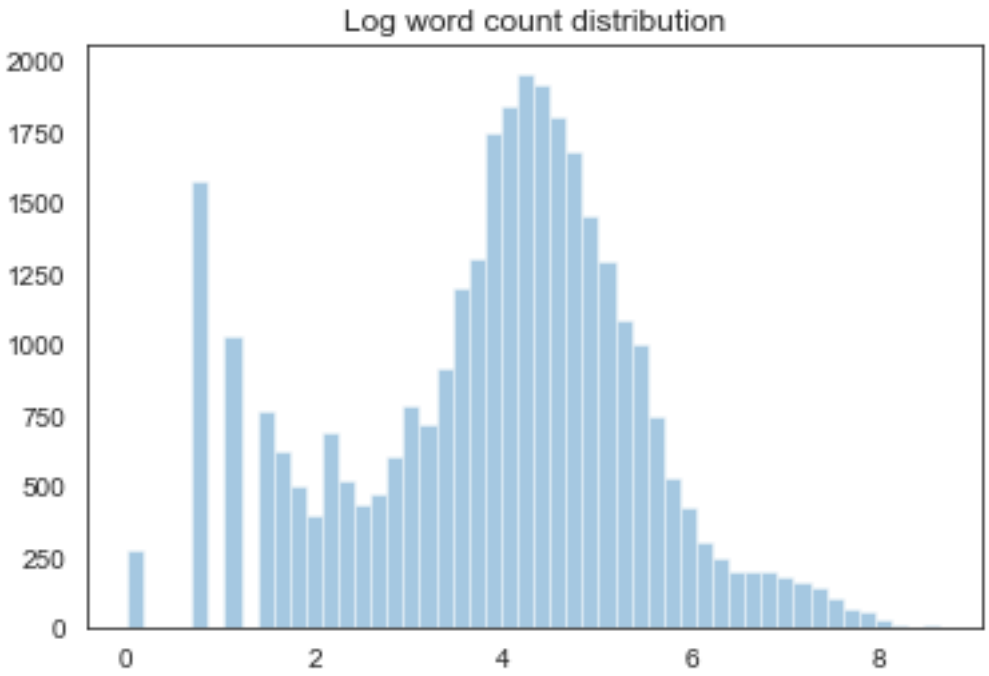
[8]:
```
len(documents)
```

[8]:
```
32047
```

1.3 Explore Data

1.3.1 Tokens per document

[9]:
```
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')
sns.despine();
```



[10]:
```
word_count.describe(percentiles=np.arange(.1, 1.0, .1))
```

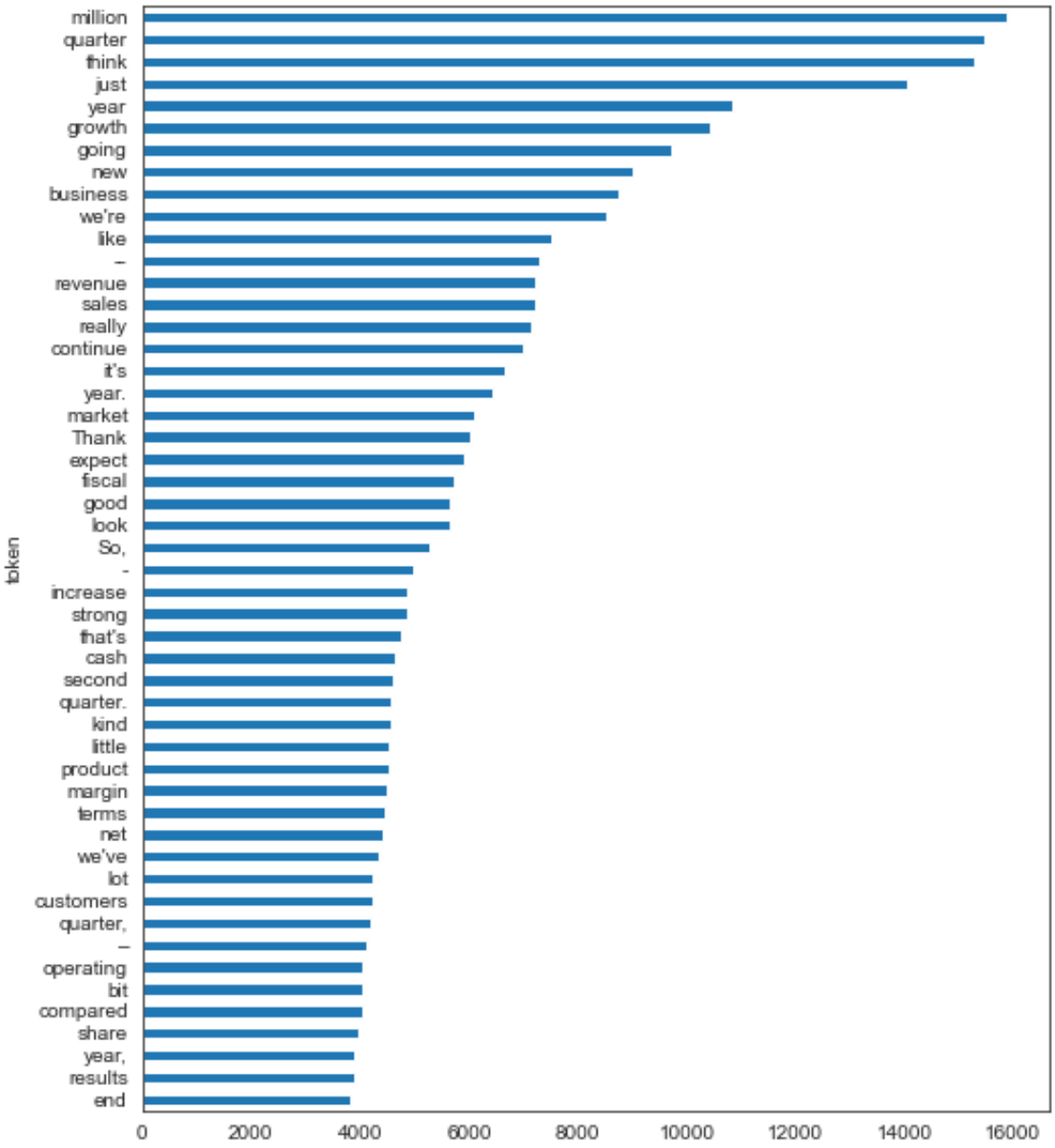
[10]:
```
count   32,047.00
mean       137.33
std        283.60
min          1.00
10%          4.00
20%         12.00
30%         28.00
40%         45.00
50%         62.00
60%         82.00
70%        111.00
80%        161.00
90%        273.00
max      5,718.00
dtype: float64
```

[11]:
```
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 30000
```

1.3.2 Most frequent tokens

[12]:
```
(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)))
sns.despine()
plt.tight_layout();
```



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.

[13]:
```
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)
```

[14]:
```
nlp = spacy.load('en')
iter_docs = (doc for doc in documents)
clean_docs = []
for i, document in enumerate(nlp.pipe(iter_docs, batch_size=100, n_process=8),␣
↪1):
    if i % 1000 == 0:
        print(f'{i/len(documents):.2%}', end=' ', flush=True)
    clean_docs.append(clean_doc(document))
```

```
3.12% 6.24% 9.36% 12.48% 15.60% 18.72% 21.84% 24.96% 28.08% 31.20% 34.32% 37.45%
40.57% 43.69% 46.81% 49.93% 53.05% 56.17% 59.29% 62.41% 65.53% 68.65% 71.77%
74.89% 78.01% 81.13% 84.25% 87.37% 90.49% 93.61% 96.73% 99.85%
```

[16]:
```
results_path = Path('results', 'earnings_calls')
if not results_path.exists():
    results_path.mkdir()
```

[17]:
```
clean_text = results_path / 'clean_text.txt'
```

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

[18]:
```
14079406
```

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]
```

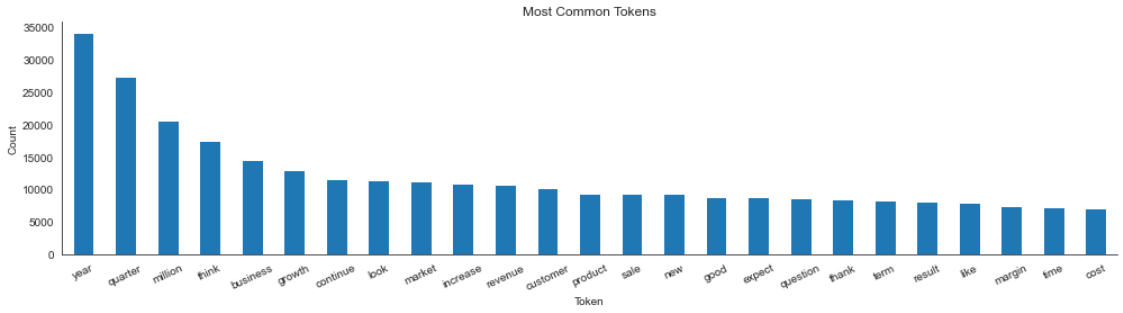
```
    if len(line) > 10:
        docs.append(' '.join(line))

len(docs)
```

[19]: 22571

[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 20000

[21]:
```
ax = (token_count.set_index('token').squeeze().iloc[:25].sort_values(
    ascending=False).plot.bar(figsize=(14, 4), rot=25, title='Most Common␣
↪Tokens'))
ax.set_ylabel('Count')
ax.set_xlabel('Token')
sns.despine()
plt.gcf().tight_layout()
```



[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)

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

[25]:
```
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)
    sns.despine()
    fig.tight_layout()
    if save:
        fig.savefig(results_path / 'earnings_call_wordlist', dpi=300)
```

[26]:
```
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])
    sns.despine()
    fig.tight_layout();
```

[27]:
```
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.

[28]:
```
min_df = .005
max_df=.25
ngram_range=(1, 1)
binary = False
```

[29]:
```
vectorizer = CountVectorizer(stop_words=frequent_words,
                             min_df=min_df,
                             max_df=max_df,
                             ngram_range=ngram_range,
                             binary=binary)
```

[30]:
```
dtm = vectorizer.fit_transform(docs)
tokens = vectorizer.get_feature_names()
dtm.shape
```

[30]:
```
(22571, 1526)
```

[31]:
```
corpus = Sparse2Corpus(dtm, documents_columns=False)
id2word = pd.Series(tokens).to_dict()
dictionary = Dictionary.from_corpus(corpus, id2word)
```

1.6.2 Model Settings

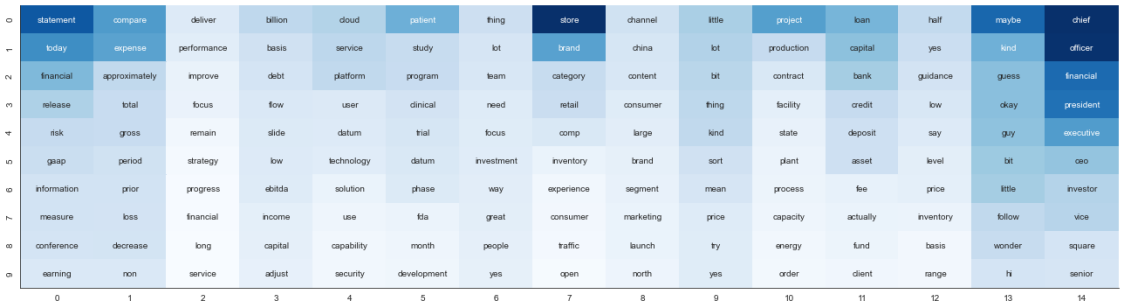
[32]:
```
num_topics=15
chunksize=50000
passes=25
update_every=None
alpha='auto'
eta='auto'
decay=0.5
offset=1.0
eval_every=None
iterations=50
```

```
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).

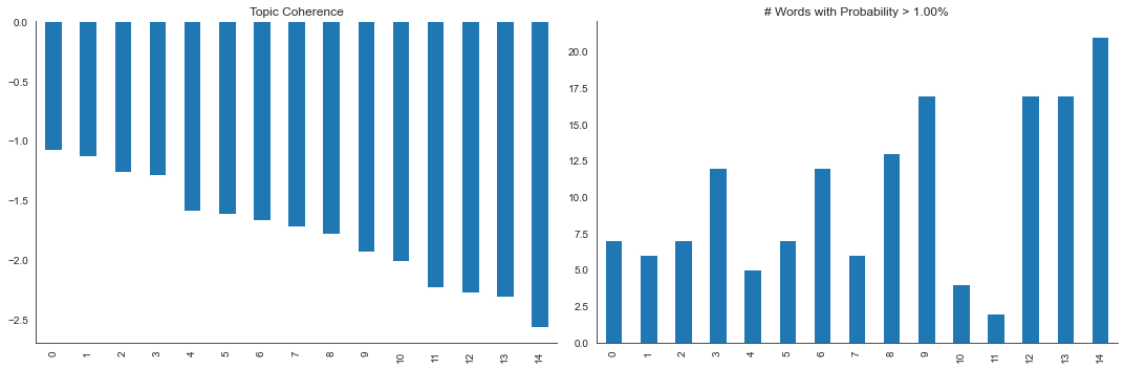
[33]:
```
lda = LdaModel(corpus=corpus,
               id2word=id2word,
               num_topics=num_topics,
               chunksize=chunksize,
               update_every=update_every,
               alpha=alpha,
               eta=eta,
               decay=decay,
               offset=offset,
               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)
```

[34]:
```
show_word_list(model=lda, corpus=corpus, save=True)
```



1.6.3 Topic Coherence

[35]:
```
show_coherence(model=lda, corpus=corpus, tokens=tokens)
```



1.6.4 pyLDAVis

[36]:
```
vis = prepare(lda, corpus, dictionary, mds='tsne')
pyLDAvis.display(vis)
```

[36]:
```
<IPython.core.display.HTML object>
```

[40]:
```
pyLDAvis.save_html(vis, (results_path / f'lda_15.html').as_posix())
```

1.6.5 Show documents most represenative of each topic

[41]:
```
show_top_docs(model=lda, corpus=corpus, docs=docs)
```

```
0 sure happy talk strategy Steve look big series announcement Dell Technologies
World handful week ago talk Dell Technologies Cloud specifically talk build
cloud platform base HCI CI stack VMware Cloud Foundation organization want
deploy prem cloud hybrid cloud consistency management automation provide tool
high level integration VxRail VCF product VMware VCF product think
differentiated offer continue build add capability talk extension platform
primary storage array continue able enable customer build edge deployment core
datum center deployment offer announce data center service basically ability
public cloud experience prem private cloud deliver service private datum center
edge network fully manage offer subscription build VxRail VMware Cloud
Foundation install manage cut like cloud service customer consume today extend
cloud strategy announcement Microsoft essentially allow customer want use public
cloud public cloud service VMware software define datum center run cloud include
AWS marketplace today announce Dell Technologies World VMware service Azure
pretty excited believe customer wide range capability multi cloud world customer
today deploy average different cloud architecture today want consistency
management automation deploy provision announce strategy allow vm application
container edge core datum center public cloud orchestrate specifically VMware
capability build highly integrate infrastructure product help customer strategy
spend vast Monday Tuesday Dell Technologies World talk help
```

```
            0    1
0       cloud 0.02
1     service 0.02
2    platform 0.02
3        user 0.01
4       datum 0.01
5  technology 0.01
6    solution 0.01
7         use 0.01
8  capability 0.01
9    security 0.01
```

1 thank good afternoon like thank time join SG Blocks 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 Blocks 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 0.05
1        today 0.04
2    financial 0.03
3      release 0.02
4         risk 0.02
5         gaap 0.02
6  information 0.02
7      measure 0.02
8   conference 0.01
9      earning 0.01
```

2 yes think problem big problem market rebate activity rebate shoot people buy year supply lense order big rebate year supply lense month month worth actual wearing right people necessarily wear lense single day lot lense kind market people shelf forth think thing market wear growth mean talk number people

contact lense wearer growth world new wearer come contact lense fantastic market
kind agree like little disappointed market growth Americas hope little bit
strong think personally CooperVision little bit half year feel good think market
end end day kind grower Asia Pac region strong look historically kind little bit
talk recession market softness mean market grow north like mean pull stat matter
fact happen handy contact lens market grow kind recession market overall grow
bounce pretty recessionary resistant lot trade forth tie fact global growth
wearer come market world outside good growth toric multifocal people fit
correctly conversion daily daily silicone hydrogel help forth pretty recession
resistant mean look pretty bad market contact lens market grow grow

	0	1
0	little	0.02
1	lot	0.02
2	bit	0.02
3	thing	0.02
4	kind	0.02
5	sort	0.02
6	mean	0.01
7	price	0.01
8	try	0.01
9	yes	0.01

3 thank Ramon good morning comment balance outlook mention release reiterate
component guidance continue expect organic revenue growth core effective tax
rate approximately core constant currency EPS decline approximately free cash
flow approximately billion total cash return shareholder approximately billion
comprise dividend approximately billion share repurchase approximately billion
organic revenue growth core constant currency EPS drop high year target imply
growth rate balance year growth rate let address think start line rate organic
revenue growth FLNA extraordinary expect rate growth continue balance year
second represent easy lap year balance year lap difficult especially half year
EPS perspective consider follow benefit lap onetime frontline bonus benefit
approximately million insurance recovery current year balance year lapping gain
strategic asset sale franchise gain insurance recovery notably rate net
commodity inflation include impact transactional foreign exchange expect
accelerate second quarter finally pace plan reinvestment accelerate course year
reflect core EPS operate margin performance open question operator question

	0	1
0	billion	0.02
1	basis	0.02
2	debt	0.01
3	flow	0.01
4	slide	0.01
5	low	0.01
6	ebitda	0.01
7	income	0.01
8	capital	0.01
9	adjust	0.01

4 thank Greg turn oncology study Phase clinical study design open label study

explore safety tolerability efficacy follow intra tumoral injection intend produce single strand antisense RNA direct inhibit expression Epidermal Growth Factor Receptor EGFR protein stimulate induce cell differentiation proliferation know significantly amplify cancer cell administer target lesion week week patient squamous cell carcinoma head neck know HNSCC patient refractory standard therapy surgery chemotherapy immunotherapy continue good progress start study site Chris Lifehouse active hope site active patient enrol near future addition regulatory approval Ministry Health Russia hope russian site active month June total anticipate clinical site Australia Russia primary outcome trial measure size size lesion technique CT MRI positive response treatment reduce overall size tumor completely ablate lesion additionally look secondary endpoint progression free survival overall survival duration response disease control rate safety tolerability addition monitor molecular marker biobsolete lesion monitor presence antisense RNA produce vector result impact egfr level study stage design allow stop study base futility success end stage patient enrol monitor primary outcome measure assume enrollment rate expect interim analysis anticipate occur end calendar year touch stand regulatory clinical perspective program treat oculopharyngeal muscular dystrophy OPMD develop therapeutic treat dysphagia associate OPMD disease symptom OPMD varied depend severity rate progression pathophysiology current symptom include ptosis dysphagia leg weakness dysphagia impair swallow function cause majority health problem patient OPMD disease progress incidence hospitalization aspiration seriousness result lung infection significantly decrease quality life life threaten addition malnutrition factor equation currently write clinical study patient clinically genetically diagnose OPMD impairment swallow function patient receive single intra muscular dose cricopharyngeal muscle throat clinical protocol design dose escalation study enroll patient increase dose time period cohort ensure stay proceed dose reach dose define dose maximally effective dose enroll additional number patient patient follow year study primary endpoint safety tolerability patient look quantitative clinical improvement swallow patient report improvement swallowing quality life meet month face face OPMD Clinical Advisory Board update clinical plan finalize clinical protocol design group comprise key opinion leader doctor help manage treatment option OPMD patient separate group expert quantitative assessment swallow input design clinical study way hope maximize opportunity success positive time forward approval turn regulatory advance program important especially gene therapy product novel silence replace approach discuss IND enable study clinical plan regulatory agency improve likelihood pass regulatory requirement enter clinic mind past quarter meet discuss plan regulatory agency key OPMD country Canada quarter France development plan receive high level enthusiasm ingenuity single vector approach silence disease cause gene simultaneously express novel healthy copy gene feedback meeting incorporate clinical study design ongoing ind enable work feel position progress clinic pathway set time complete required toxicology manufacturing work support high quality regulatory filing turn David Suhy ongoing toxicology manufacturing work

	0	1
0	patient	0.03
1	study	0.02

```
2      program 0.02
3     clinical 0.02
4        trial 0.01
5        datum 0.01
6        phase 0.01
7          fda 0.01
8        month 0.01
9  development 0.01
5 think look share balance product category man woman brand activation test
membership event activity able leverage improved datum analytic digital
marketing mean guest respond share store traffic plus e commerce good healthy
metric highly engage guest typically business significant swing week season
season holiday holiday pleased momentum consistent traffic drive big piece
business new guest exist guest balance product range
            0    1
0       store 0.05
1       brand 0.03
2    category 0.02
3      retail 0.02
4        comp 0.01
5   inventory 0.01
6  experience 0.01
7    consumer 0.01
8     traffic 0.01
9        open 0.01
6 somewhat encouraged opportunity little big opportunity come way hope continue
far talent talent incredibly important company like retain good talent add
evolve provide marketplace need evolve hunt good people help business
development resume background capability experience relationship key area
opportunity enormous outreach term identify specific type talent quality skill
need folk talent recruitment effort board important company look generation
leadership company develop talent play leadership role company change forward
grow evolve need cultivate talent internally bring people opportunity forward
carrier carrier goal super important attract talent outside far create
additional prospect create generation leadership company interested people come
organization involve path want term growth merger acquisition forth bring folk
successful thing maybe differently different method different kind idea
different approach solve problem develop new offering market type people
interesting talent internal develop internally attract outside care people
people super important care people convince stick continue career successful
            0    1
0       thing 0.02
1         lot 0.01
2        team 0.01
3        need 0.01
4       focus 0.01
5  investment 0.01
6         way 0.01
```

```
7       great 0.01
8      people 0.01
9         yes 0.01
7 okay kind leveling deposit pressure March probably characterize pretty
difficult deposit environment January February come increase relate steepness
curve Steve remember invest loan short end curve steepness matter Fund Banking
division Signature Financial tie short end curve certainly traditional asset
base lending group tie short end curve area growth Venture Banking Group come
short end curve steepness matter
          0    1
0      loan 0.03
1   capital 0.03
2      bank 0.02
3    credit 0.02
4   deposit 0.02
5     asset 0.02
6       fee 0.01
7  actually 0.01
8      fund 0.01
9    client 0.01
8 thank Ed thank join today KLA Tencor deliver outstanding result September
quarter drive company technology market leadership compelling value diversified
product service portfolio enable customer success September quarter result
company benefit strategy market leadership revenue diversification furthermore
demonstrate critical nature process control enable technology inflection
semiconductor industry include growth market vertical NAND EUV adoption China
semiconductor industry expect closing quarter Orbotech acquisition KLA Tencor
extend market reach electronics value chain enhance company product service
portfolio core wfe address new growth market technological complexity rise KLA
extend core competency exciting new opportunity perspective current industry
demand environment notwithstanding recent adjustment capacity investment memory
customer delay logic spend second half believe long term factor underpin demand
wafer fab equipment industry remain sound result key element include diversified
end market semiconductor ongoing commitment customer drive innovation leading
edge technology roadmap discipline market drive capacity planning high level
investment require address increase design complexity advanced device
architecture expect industry driver persist deliver long term growth value
creation opportunity company industry stakeholder recent highlight demonstrate
successful execution company strategic focus technology market leadership market
leader process control KLA Tencor help enable growth semiconductor industry
China memory foundry manufacturer region invest high level accelerate process
development yield learn expect momentum China continue customer progress
technology roadmap ramp new production capacity initial phase multiyear
investment cycle additionally past quarter exceptional demand bare waver mask
inspection product growth mask inspection business consistent planned increase
CapEx lead foundry customer ramp capacity support high number customer tapeout
node begin aggressive development include EUV development bare wafer KLA market
lead inspection metrology tool key enabler growth wafer supply address capacity
```

expansion memory product essential help customer segment meet stringent design specification wafer flatness cleanliness finally KLA Tencor service business continue highlight forward trajectory deliver high single digit low double digit annual revenue growth low business volatility consistent historical trend long term growth service tie expansion instal base increase uptime requirement current node production trailing edge trail edge fab currently run near utilization create new opportunity product enhancement upgrade summary turn Bren KLA Tencor deliver outstanding performance September quarter drive company strong execution market leadership momentum marketplace diversified growth market critical role company product service play enable customer success position strong finish like turn Bren review financial result Bren

	0	1
0	deliver	0.01
1	performance	0.01
2	improve	0.01
3	focus	0.01
4	remain	0.01
5	strategy	0.01
6	progress	0.01
7	financial	0.01
8	long	0.01
9	service	0.01

9 couple month ago look trade war think affect market think affect small size decision start fix vessel come free quarter year bit second quarter right fix number vessel actually probably vessel come free depend thing continue charter certain period small vessel actually instal scrubber instal scrubber big vessel Newcastlemaxes Capes January mean dedicated big vessel small vessel reason flexibility fix spot market low vessel open income short period probably short period low spot rate recover continue similar policy

	0	1
0	half	0.02
1	yes	0.02
2	guidance	0.02
3	low	0.01
4	say	0.01
5	level	0.01
6	price	0.01
7	inventory	0.01
8	basis	0.01
9	range	0.01

10 objective primary production cost cost quarter primary produce production roughly purify produce month January primary quality brine deteriorate quarter impact negatively low brine quality require large energy plant brine plant large reagent expect able stabilize brine cost close range think achievable time effort require stabilization brine quality

	0	1
0	project	0.03
1	production	0.02

```
2    contract 0.01
3    facility 0.01
4       state 0.01
5       plant 0.01
6     process 0.01
7    capacity 0.01
8      energy 0.01
9       order 0.01
11 thing work diligently fact learning mvmt brand level content quality content
produce Movado drive immediate result pleased business second half year continue
strong momentum strong momentum business believe able talk consumer different
way ampe level content use reach consumer reach change dramatically direct
relationship know customer rewarding reach vehicle reach build funnel digital
perspective continue build brand awareness image brand vitally important
           0    1
0    channel 0.02
1      china 0.01
2    content 0.01
3   consumer 0.01
4      large 0.01
5      brand 0.01
6    segment 0.01
7  marketing 0.01
8     launch 0.01
9      north 0.01
12 Tracy Ward Investor Relations Tom Olinger Chief Financial Officer Hamid
Moghadam Chairman CEO Gary Anderson Chief Executive Officer Europe Asia Chris
Caton Senior Vice President Global Head Research Mike Curless Chief Investment
Officer Ed Nekritz Chief Legal Officer Gene Reilly Chief Executive Officer
Americas Colleen McKeown Chief Human Resources Officer
           0    1
0      chief 0.05
1    officer 0.05
2  financial 0.04
3  president 0.04
4  executive 0.03
5        ceo 0.03
6   investor 0.02
7       vice 0.02
8     square 0.01
9     senior 0.01
13 okay fair want ask question kind cash flow cash flow cycle able expect year
know inventory build little bit hope improve result term thinking kind cash flow
cycle remainder year look kind maybe little drag expect time think Marc
remainder year couple quarter kind hold serve cash flow operation kind nice cash
flow context
        0    1
0   maybe 0.04
```

```
1    kind 0.03
2   guess 0.03
3    okay 0.03
4     guy 0.03
5     bit 0.02
6  little 0.02
7  follow 0.02
8  wonder 0.02
9      hi 0.01
```

14 thank Wahid good afternoon AeroVironment fiscal fourth quarter result follow revenue continue operation fourth quarter fiscal million decrease million fourth quarter fiscal revenue million decrease decrease product delivery million decrease service revenue fourth quarter fiscal revenue major product program follow small UAS million TMS million HAPS million million Gross margin continue operation fourth quarter fiscal million revenue compare million revenue fourth quarter fiscal decrease gross margin primarily decrease product margin million decrease service margin million gross margin percentage revenue decrease primarily increase proportion service revenue total revenue unfavorable service revenue mix high CIS inventory reserve look rest income statement expense continue operation fourth quarter fiscal million revenue compare expense million revenue fourth quarter fiscal increase primarily CIS fix asset impairment charge rate adoption Quantix AV DSS solution slow expect fourth quarter lower future outlook unit sale result low forecast impairment charge million s Quantex AV DSS fix asset expense continue operation fourth quarter fiscal million revenue compare expense million revenue fourth quarter fiscal income continue operation fourth quarter fiscal revenue compare million fourth quarter fiscal decrease income operation primarily decrease gross margin million increase expense million increase expense million net income fourth quarter fiscal million compare net income million fourth quarter fiscal increase net income income transition service agreement buyer EES business high income investment effective income tax rate continue operation minus fourth quarter fiscal compare effective income tax rate fourth quarter fiscal decrease effective tax rate fourth quarter fiscal reduction fiscal federal statutory rate low pre tax profit equity method investment activity net tax fourth quarter fiscal loss million dilute share compare loss million net tax fourth quarter fiscal net income continue operation attributable AeroVironment fourth quarter fiscal million dilute share compare net income continue operation attributable AeroVironment million cent diluted share fourth quarter fiscal net loss discontinue operation net tax fourth quarter fiscal million compare loss discontinue operation net tax million fourth quarter fiscal year fiscal result revenue fiscal million increase million compare million fiscal increase revenue increase service revenue million increase product revenue million inception date revenue hapsmobile million total value contract hapsmobile million consist million design development agreement million preliminary design related effort million remain contract include portion currently unfunded gross margin fiscal million revenue compare million fiscal increase increase product margin million increase service margin million Gross margin percentage revenue increase primarily favorable product mix partially offset unfavorable service mix increase CIS inventory reserve charge

expense fiscal million revenue compare expense million revenue fiscal increase primarily million excessive impairment charge CIS business expense relate transition service agreement buyer ees business expense fiscal million revenue compare expense million revenue fiscal net income fiscal million compare prior year net income million net income increase primarily litigation settlement income earn transition service agreement buyer EES business increase income effective income tax rate continue operation fiscal compare effective income tax rate fiscal effective income tax rate fiscal include impact time defer tax expense result measurement exist defer tax asset liability million decrease effective income tax rate reduction fiscal federal statutory rate equity method investment activity net tax fiscal loss million dilute share compare loss million net tax fiscal increased loss increase ownership hapsmobile joint venture high investment hapsmobile joint venture net income continue operation attributable AeroVironment fiscal million dilute share compare million dilute share fiscal net income discontinue operation net tax fiscal million dilute share compare loss discontinue operation net tax million fiscal dilute share fiscal include million gain net tax sale ees business funded backlog April million decrease fourth quarter fiscal increase million quarter fiscal backlog million turn balance sheet cash cash equivalent investment end fourth quarter fiscal total million increase million end fiscal cash cash equivalent investment million net account receivable include unbilled receivable retention end fourth quarter fiscal total million unbilled receivables retention balance million inclusive million related party total day sale outstanding continue operation fourth quarter fiscal year approximately day compare day fourth quarter fiscal year net inventory end fourth quarter fiscal year million compare million end fourth quarter fiscal year day inventory outstanding fourth quarter fiscal year approximately day compare day fourth quarter fiscal year account payable end fourth quarter fiscal year million compare million end fourth quarter fiscal year total day payable outstanding fourth quarter fiscal year approximately compare day fourth quarter fiscal year turn capital expenditure fourth quarter fiscal year invest approximately million property improvement capital equipment continue operation recognize million depreciation amortization expense update fiscal visibility today fourth quarter end backlog expect execute fiscal million quarter date booking anticipate execute fiscal million unfunded backlog incrementally fund contract anticipate recognize revenue balance year million add million fiscal year midpoint revenue guidance range anticipate year effective tax rate range like turn thing Wahid

	0	1
0	compare	0.03
1	expense	0.03
2	approximately	0.02
3	total	0.02
4	gross	0.02
5	period	0.02
6	prior	0.02
7	loss	0.02
8	decrease	0.01
9	non	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.

[43]:
```
with pd.HDFStore(results_path / 'results.h5') as store:
    perplexity = store.get('perplexity')
    coherence = store.get('coherence')
```

[44]:
```
perplexity.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 496 entries, 0 to 15
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   vocab_size  496 non-null    int64
 1   test_vocab  496 non-null    int64
 2   min_df      496 non-null    int64
 3   max_df      496 non-null    float64
 4   binary      496 non-null    bool
 5   num_topics  496 non-null    int64
 6   passes      496 non-null    int64
 7   perplexity  496 non-null    float64
dtypes: bool(1), float64(2), int64(5)
memory usage: 31.5 KB
```

[45]:
```
coherence.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8370 entries, 0 to 713
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   topic       8370 non-null   int64
 1   passes      8370 non-null   object
 2   num_topics  8370 non-null   object
 3   coherence   8370 non-null   float64
 4   min_df      8370 non-null   int64
```

```
 5   max_df      8370 non-null   float64
 6   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

[46]:
```
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:                Sat, 20 Jun 2020   Prob (F-statistic):          5.53e-116
Time:                        16:35:51   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

[47]:
```
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:                Sat, 20 Jun 2020   Prob (F-statistic):               0.00
Time:                        16:35:55   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]
------------------------------------------------------------------------------
```

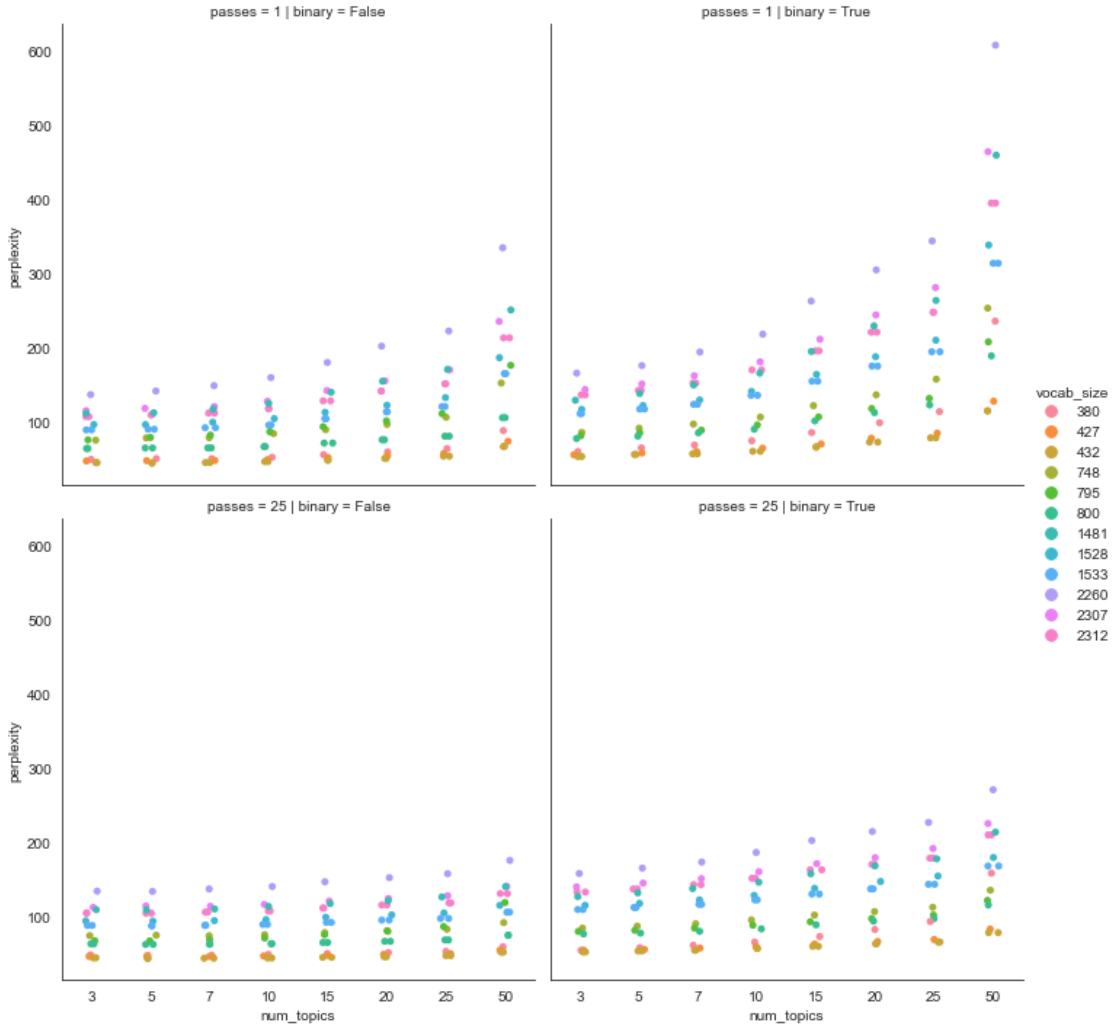
```
-
const          -1.5492      0.023    -66.490      0.000      -1.595
-1.504
topic_1        -0.1076      0.020     -5.501      0.000      -0.146
-0.069
topic_2        -0.2316      0.020    -11.553      0.000      -0.271
-0.192
topic_3        -0.3080      0.019    -16.228      0.000      -0.345
-0.271
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

[48]:
```
sns.catplot(x='num_topics',
            y='perplexity',
            data=perplexity,
            hue='vocab_size',
            col='binary',
            row='passes',
            kind='strip');
```

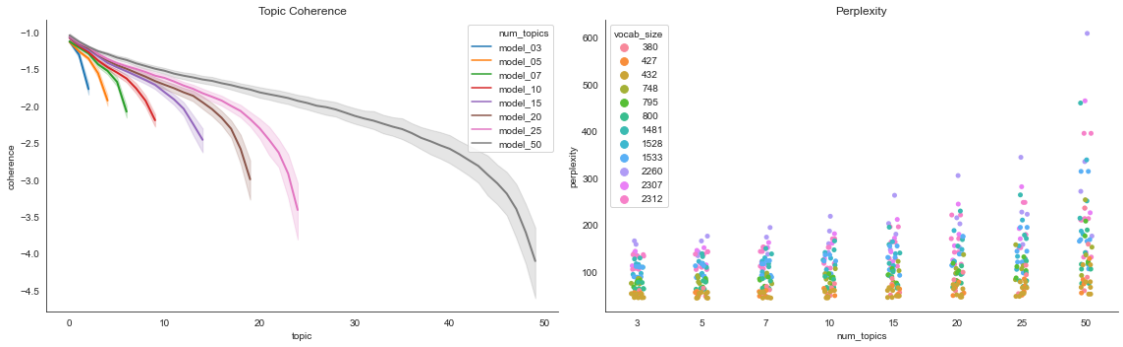


[49]:
```
coherence.num_topics = coherence.num_topics.apply(lambda x: f'model_{int(x):
↪0>2}')
perplexity.min_df = perplexity.min_df.apply(lambda x: f'min_df_{int(x):0>3}')
```

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.

[50]:
```
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')
sns.despine()
fig.tight_layout();
```



[ ]:
```

```