In [10]:

```
import os
import pandas as pd
pd.set_option('display.max_colwidth', -1) #to prevent cell display truncation

import re

from rake_nltk import Metric, Rake #for RAKE scored keyword phrases
from gensim.summarization.summarizer import summarize #for TextRank scored sentences
```

Summarization techniques used:

Rapid Automatic Keyword Extraction (RAKE)

RAKE has been shown to perform better than Textrank in keywords extraction. This technique splits given text into candidate keywords, wherever stop words or phrase delimiters occur. Then it tabulates the frequency of co-ocurrence of entire candidate keywords and their member words. These frequencies are used to create a RAKE score using a simple formula. Importantly, there is a preference for words that occur in longer phrases/ candidate keywords.

TextRank

Extract text units (sentences) and use them as vertices in a graph. Build edges representing similarity/ overlap/ common tokens between the vertices. Through a recursive process, score each vertex by the number of edges connected to it. Such scoring also accounts for the number of edges connected to the vertices voting for the current vertex.

In [11]:

```
#function that summarizes text
def summarizer (caseStudies, topRake, TR):
    RAKEkeyWords = []
    textRankSentSumm = []
    r = Rake(ranking_metric=Metric.WORD_DEGREE)
    reviewFiles = []
    for row in range(len(caseStudies)):
        caseStudy = caseStudies.loc[row,'Content']
        #print(caseStudies.loc[row,'fileNames'])
        #RAKE scores calculation
        r.extract_keywords_from_text(caseStudy)#does the heavy lifting of calculating RAKE scores
        keyWordEach = r.get_ranked_phrases()[0:topRake]#limits to the top x phrases, by RAKE score
        keyWordEachCons = '| '.join(keyWordEach)
        RAKEkeyWords.append(keyWordEachCons)
        #textRank scores calculation
        try:
            textRankSentSummRow = summarize(caseStudy, word_count=TR)#textRank scoring and limiting the number of results returned
            textRankSentSummRowCons = ''.join(textRankSentSummRow)
            if not textRankSentSummRowCons:
                textRankSentSummRowCons = "<>"#filler when no sentences are shortlisted by textRank
            textRankSentSumm.append(textRankSentSummRowCons)
        except Exception:
            reviewFiles.append(caseStudies.loc[row,'fileNames'])
            textRankSentSummRowCons = caseStudy#header only/ single sentence files
            textRankSentSumm.append(textRankSentSummRowCons)
            pass

    RAKEkeyWords = pd.DataFrame(RAKEkeyWords)
    textRankSentSumm = pd.DataFrame(textRankSentSumm)
    caseStudieswSumm = pd.concat([RAKEkeyWords, textRankSentSumm], axis=1)
    caseStudieswSumm = pd.concat([RAKEkeyWords, textRankSentSumm, caseStudies[['fileNames', 'Content']]], axis=1)
    caseStudieswSumm.columns = ['RAKEkeyWords', 'textRankSentSumm', 'fileName', 'Content']
    return caseStudieswSumm[['RAKEkeyWords', 'textRankSentSumm', 'fileName']]
```

In [12]:

```
temp = open("/media/sf_ForVirtualBox/KM/common/caseStudiesPPTX.csv", 'r', encoding = 'latin-1')
caseStudiesPPTX = pd.read_csv(temp)
caseStudies = caseStudiesPPTX.iloc[[13, 14, 59, 121, 225], :]#randomly selecting .pptx
caseStudies = caseStudies[['fileNames', 'Content']]
caseStudies.reset_index(drop=True, inplace=True)
print('List of sample .pptx', '\n' ,caseStudies['fileNames'])
```

```
List of sample .pptx
 0    UBS_DB_Migration _Case Study.pptx
1    UHG Upgrade Response - v3.9.2.pptx
2    Sybase to Oracle Migration â□□ Leading US Healthcare Clearing House.pptx
3    Pruitt Health_PeopleSoft HCM  FSCM Managed Services Support.pptx
4    Oshkosh_Manlog_Oracle EBS iSupplier_R12.0.4_Implementation.pptx
Name: fileNames, dtype: object
```

In [13]:

```
#Trial 1
summarizer(caseStudies, 1, 10)
```

Out[13]:

	RAKEkeyWords	textRankSentSumm	fileName
0	oracle staging tables .. execute oracle scripts	<>	UBS_DB_Migration _Case Study.pptx
1	create weblogic domainsinstall oracle soa 12c components â □□ soa	<>	UHG Upgrade Response - v3.9.2.pptx
2	oracle migration â □□ leading us healthcare clearing house	<>	Sybase to Oracle Migration â□□ Leading US Healthcare Clearing House.pptx
3	sql server bids 2008 r2ms sql server 2008 r2sharepoint 2010	. peoplesoft hcm & fscm managed services support for a major private us healthcare providers in southeast (pruitt health).	Pruitt Health_PeopleSoft HCM FSCM Managed Services Support.pptx
4	oracle adaptersoracle soa components like bpel	<>	Oshkosh_Manlog_Oracle EBS iSupplier_R12.0.4_Implementation.pptx

In [14]:

```
#Trial 2
summarizer(caseStudies, 3, 20)
```

Out[14]:

	RAKEkeyWords	textRankSentSumm	fileName
0	oracle staging tables .. execute oracle scripts| csv data file .. column level mapping document| every column value .. execute control files using	<>	UBS_DB_Migration _Case Study.pptx
1	create weblogic domainsinstall oracle soa 12c components â □□ soa| soa governance leveraging oracle api managerstreamline api editing| implement soa governance leveraging oracle api manager	<>	UHG Upgrade Response - v3.9.2.pptx
2	oracle migration â □□ leading us healthcare clearing house| overall effort reductionteam compositiononsite offshore ratio â □□ 20 fte| oracle new oracle database support	<>	Sybase to Oracle Migration â□□ Leading US Healthcare Clearing House.pptx
3	sql server bids 2008 r2ms sql server 2008 r2sharepoint 2010| long term care automationmass merit increase process paycheck reconciliation process| biztalk server 2013 r2sql server 2014sis	. peoplesoft hcm & fscm managed services support for a major private us healthcare providers in southeast (pruitt health).	Pruitt Health_PeopleSoft HCM FSCM Managed Services Support.pptx
4	oracle adaptersoracle soa components like bpel| oracle b2b applications using| oracle ebs isupplier implementation services	<>	Oshkosh_Manlog_Oracle EBS iSupplier_R12.0.4_Implementation.pptx

In [15]:

```
#Trial 3
summarizer(caseStudies, 5, 30)
```

Out[15]:

	RAKEkeyWords	textRankSentSumm	fileName
0	oracle staging tables .. execute oracle scripts| csv data file .. column level mapping document| every column value .. execute control files using| underlying database ms access .- entire data set| main business tables using various oracle sub programs	<>	UBS_DB_Migration _Case Study.pptx
1	create weblogic domainsinstall oracle soa 12c components â □□ soa| soa governance leveraging oracle api managerstreamline api editing| implement soa governance leveraging oracle api manager| enable soa governance using oracle api managergo| complex engagements oracle soa upgrade 11g	<>	UHG Upgrade Response - v3.9.2.pptx
2	oracle migration â □□ leading us healthcare clearing house| overall effort reductionteam compositiononsite offshore ratio â □□ 20 fte| oracle new oracle database support| new oracle database contains 2 node rac| acceleratorsoracle work bench tool â □□ 30	<>	Sybase to Oracle Migration â□□ Leading US Healthcare Clearing House.pptx
3	sql server bids 2008 r2ms sql server 2008 r2sharepoint 2010| long term care automationmass merit increase process paycheck reconciliation process| biztalk server 2013 r2sql server 2014sis| actcasamba system integrationattendance enterprise integrationcreated new sharepoint service application| expense reportcms pbj interface biztalk â □□ casamba	. peoplesoft hcm & fscm managed services support for a major private us healthcare providers in southeast (pruitt health).	Pruitt Health_PeopleSoft HCM FSCM Managed Services Support.pptx
4	oracle adaptersoracle soa components like bpel| oracle b2b applications using| oracle ebs isupplier implementation services| oracle application using aq| oracle soa suite	cognizant was selected as the partner to assist client in evaluating the product of choice and implemented the samecognizant utilized its proprietary fabulous framework and aimazon methodologyextensively involved in architecting, design and development of key architectural extension componentsimplemented a very effective architecture in connecting oracle b2b to oracle soa suite that would meet the enterprise needs.	Oshkosh_Manlog_Oracle EBS iSupplier_R12.0.4_Implementation.pptx