

# Text Summarization

<b>Class</b>	Algorithmic Methods of Data Mining
<b>Program</b>	M. Sc. Data Science
<b>University</b>	Sapienza University of Rome
<b>Semester</b>	Fall 2015
<b>Lecturer</b>	Carlos Castillo <a href="http://chato.cl/">http://chato.cl/</a>

## Sources:

- Lloret et al. “[Text Summarization: an Overview](#)” 2008.
- Erkan et al. “[LexRank ...](#)” 2004.
- Document summarization [slides](#).

# Single-document summary (152,621 words → 812 words)

Nine years after the start of the Trojan War, the Greek (“Achaean”) army sacks Chryse, a town allied with Troy. During the battle, the Achaeans capture a pair of beautiful maidens, Chryseis and Briseis. Agamemnon, the leader of the Achaean forces, takes Chryseis as his prize, and Achilles, the Achaeans’ greatest warrior, claims Briseis. Chryseis’s father, Chryses, who serves as a priest of the god Apollo, offers an enormous ransom in return for his daughter, but Agamemnon refuses to give Chryseis back. Chryses then prays to Apollo, who sends a plague upon the Achaean camp.

After many Achaeans die, Agamemnon consults the prophet Calchas to determine the cause of the plague. When he learns that Chryseis is the cause, he reluctantly gives her up but then demands Briseis from Achilles as compensation. Furious at this insult, Achilles returns to his tent in the army camp and refuses to fight in the war any longer. He vengefully yearns to see the Achaeans destroyed and asks his mother, the sea-nymph Thetis, to enlist the services of Zeus, king of the gods, toward this end. The Trojan and Achaean sides have declared a cease-fire with each other, but now the Trojans breach the treaty and Zeus comes to their aid.

With Zeus supporting the Trojans and Achilles refusing to fight, the Achaeans suffer great losses. Several days of fierce conflict ensue, including duels between Paris and Menelaus and between Hector and Ajax. The Achaeans make no progress; even the heroism of the great Achaean warrior Diomedes proves fruitless. The Trojans push the Achaeans back, forcing them to take refuge behind the ramparts that protect their ships. The Achaeans begin to nurture some hope for the future when a nighttime reconnaissance mission by Diomedes and Odysseus yields information about the Trojans’ plans, but the next day brings disaster. Several Achaean commanders become wounded, and the Trojans break through the Achaean ramparts. They advance all the way up to the boundary of the Achaean camp and set fire to one of the ships. Defeat seems imminent, because without the ships, the army will be stranded at Troy and almost certainly destroyed.

Concerned for his comrades but still too proud to help them himself, Achilles agrees to a plan proposed by Nestor that will allow his beloved friend Patroclus to take his place in battle, wearing his armor. Patroclus is a fine warrior, and his presence on the battlefield helps the Achaeans push the Trojans away from the ships and back to the city walls. But the counterattack soon falters. Apollo knocks Patroclus’s armor to the ground, and Hector slays him. Fighting then breaks out as both sides try to lay claim to the body and armor. Hector ends up with the armor, but the Achaeans, thanks to a courageous effort by Menelaus and others, manage to bring the body back to their camp. When Achilles discovers that Hector has killed Patroclus, he fills with such grief and rage that he agrees to reconcile with Agamemnon and rejoin the battle. Thetis goes to Mount Olympus and persuades the god Hephaestus to forge Achilles a new suit of armor, which she presents to him the next morning. Achilles then rides out to battle at the head of the Achaean army.

Meanwhile, Hector, not expecting Achilles to rejoin the battle, has ordered his men to camp outside the walls of Troy. But when the Trojan army glimpses Achilles, it flees in terror back behind the city walls. Achilles cuts down every Trojan he sees. Strengthened by his rage, he even fights the god of the river Xanthus, who is angered that Achilles has caused so many corpses to fall into his streams. Finally, Achilles confronts Hector outside the walls of Troy. Ashamed at the poor advice that he gave his comrades, Hector refuses to flee inside the city with them. Achilles chases him around the city’s periphery three times, but the goddess Athena finally tricks Hector into turning around and fighting Achilles. In a dramatic duel, Achilles kills Hector. He then lashes the body to the back of his chariot and drags it across the battlefield to the Achaean camp. Upon Achilles’ arrival, the triumphant Achaeans celebrate Patroclus’s funeral with a long series of athletic games in his honor. Each day for the next nine days, Achilles drags Hector’s body in circles around Patroclus’s funeral bier.

At last, the gods agree that Hector deserves a proper burial. Zeus sends the god Hermes to escort King Priam, Hector’s father and the ruler of Troy, into the Achaean camp. Priam tearfully pleads with Achilles to take pity on a father bereft of his son and return Hector’s body. He invokes the memory of Achilles’ own father, Peleus. Deeply moved, Achilles finally relents and returns Hector’s corpse to the Trojans. Both sides agree to a temporary truce, and Hector receives a hero’s funeral.

# Multi-document summary

## Competing visions: Clinton, Sanders spar on Mideast, economics

Democrats expressed growing skepticism Wednesday that Vice President Joe Biden could find a foothold if he entered the presidential campaign as Hillary Clinton's commanding performance in the first Democratic debate abruptly quieted murmurs about her candidacy. ([article 3](#)) Can a single night in the desert change the pace and structure of the Democrats' race to succeed President Barack Obama as president? The Hillary Clinton campaign and the party regulars who back her should hope as much. ([article 4](#)) Her challengers had to bet big to win big, and Vegas again proved to be a town built on the backs of busted gamblers. ([article 4](#)) They also sought to cast the GOP as a party focused on sowing division and denigrating minorities and women. ([article 5](#)) Far from the reality-show slug-fest of the crowded Republican debates, the Democrats' first presidential match-up offered a textbook example of what a candidates forum can do: provide a clear-eyed discussion of thorny issues and produce winners and losers. ([article 1](#)) Clinton seemed to soften her rough edges while firming up the shifting positions that have left her sounding more like a poll-tested candidate than a seasoned pro with strong convictions. ([article 1](#)) Even Democrats who are worried about her general election prospects - a growing population - dismiss the possibility of Vermont Sen. Bernie Sanders winning the nomination. ([article 2](#))

## Source articles

1. [Six take-aways from the Democratic debate](#) (L.A. Times, 10/13/2015, 737 words)
2. [Albert Hunt: Odds favor even a wounded Clinton to win](#) (dallasnews.com, 10/13/2015, 676 words)
3. [Clinton's performance at debate is silencing talk of Biden run](#) (dallasnews.com, 10/14/2015, 951 words)
4. [Editorial: What happened in Vegas? Clinton reasserts herself in debate](#) (dallasnews.com, 10/14/2015, 479 words)
5. [Competing visions: Clinton, Sanders spar on Mideast, economics](#) (dallasnews.com, 10/14/2015, 610 words)


# Multi-document summary

- Cluster-level labels in cluster search

The screenshot displays the Carrot Search engine interface. At the top left is the Carrot logo, a stylized orange carrot. Below it is a navigation bar with icons for Web, Bing, News, Images, Wiki, Jobs, PubMed, and PUT. A search bar contains the word "Orange" and a "Search" button with a "More options" link. The main content area is divided into two sections: "Folders" on the left and "Top 100 results of about 678000000 for Orange" on the right. The "Folders" section lists various categories with item counts, such as "All Topics (100)", "Orange County (10)", "Wikipedia (8)", "Family (7)", "Orange Fruit (7)", "Mobiles et Internet (6)", "Orange Is the New (6)", "Annoying Orange (4)", "Companies (4)", "Orange Colour (4)", and "Orange S.A. (4)". The "Top 100 results" section shows four search results:

- 1** [Orange | Email, shop, upgrade | EE](#) [Goo, Google]  
**Orange.co.uk** has closed. But don't worry, get your **Orange** email, manage your account, upgrade to EE and more here.  
<http://ee.co.uk/orange> [Goo, Google]
- 2** [Orange \(colour\) - Wikipedia, the free encyclopedia](#) [Goo, Google]  
**Orange** is the colour between red and yellow on the spectrum of light, and in the traditional colour wheel used by painters. Its name is derived from the fruit ...  
[https://en.wikipedia.org/wiki/Orange\\_\(colour\)](https://en.wikipedia.org/wiki/Orange_(colour)) [Google]
- 3** [orange.com: Corporate Website of Orange](#) [Ask, Goo, Google]  
All the informations about the Group: press releases, news, investors, shareholders, consolidated results, candidate, innovations, networks, corporate social ...  
<http://www.orange.com/en/home> [Ask, Goo, Google]
- 4** [Orange](#) [Goo, Google]  
Strategy that emphasizes the family's role in children's ministry. Created by Reggie Joiner and the reThink Group.  
<http://www.whatisorange.org/> [Goo, Google]

# Multi-document summary



**QUALITY INN & SUITES**  
**ARTESIA** **\$60**  
★ ★ ★ ★ ☆ (156 guest ratings)  
16905 Pioneer Blvd. California, 90701  
**no gym** . **internet access** . **restaurant** .

[Rates/Availability](#)

**People Think...** [What's Buzzing?](#) [Browse Reviews](#) [Map View](#)

- + *Great room for the price (64)*
- + *Room was clean (13)*
- *Breakfast was great (5)*

**Free Breakfast.**

Overall it wasn't too bad. It is in a very safe area and close to places to eat and the highway. The free **breakfast** was really good, it included: juice, cereal, donuts, fruit, hot biscuits and gravy, and my favorite a make your own waffle ..."

[jump.priceline.com](http://jump.priceline.com)

*What is a summary?*

# What is text summarization?

- Distilling the most important information in a text to produce an abridged version
- It can be a phrase, sentence, paragraph, etc.
  - It should be shorter than the original text

# Types

- Multi-document vs Single-document
- Extractive vs Abstractive
- Generic vs Query-driven
- Informative vs Indicative



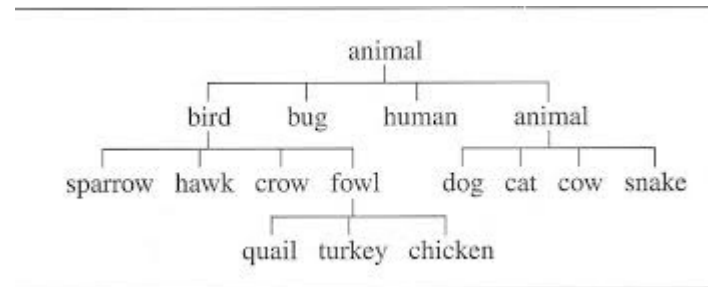
# Elements

- Words
- Position in text
  - Title or lead paragraph
  - Beginning, middle or end
- Cue words (“in conclusion, ...”)

# Connections between elements

- Proximity
- Co-occurrence
- Theasaural relationships

Hyponym and Hypernym



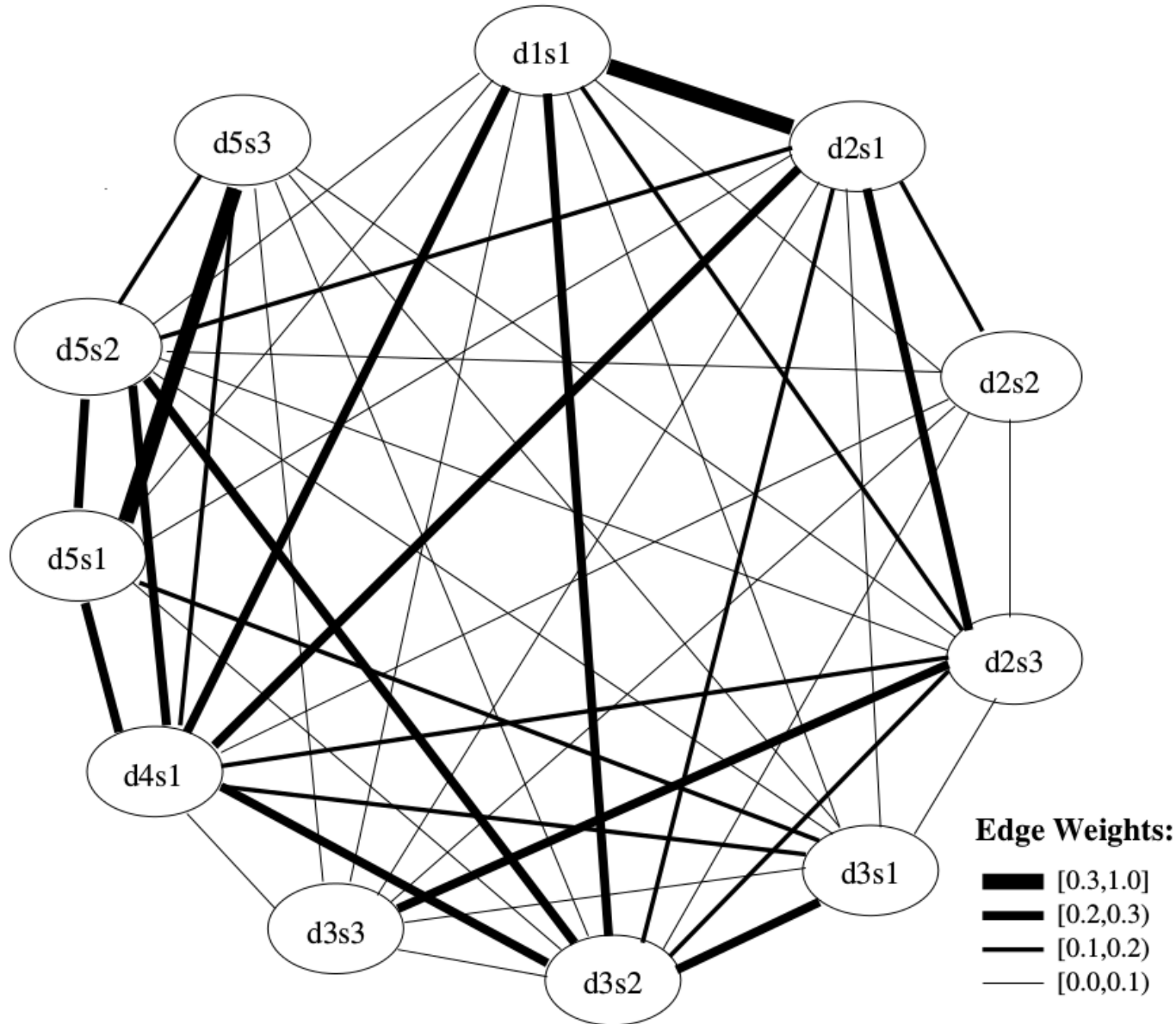
- Anaphora
  - “The city was gray; *it* was also lively”

# Key idea: centrality

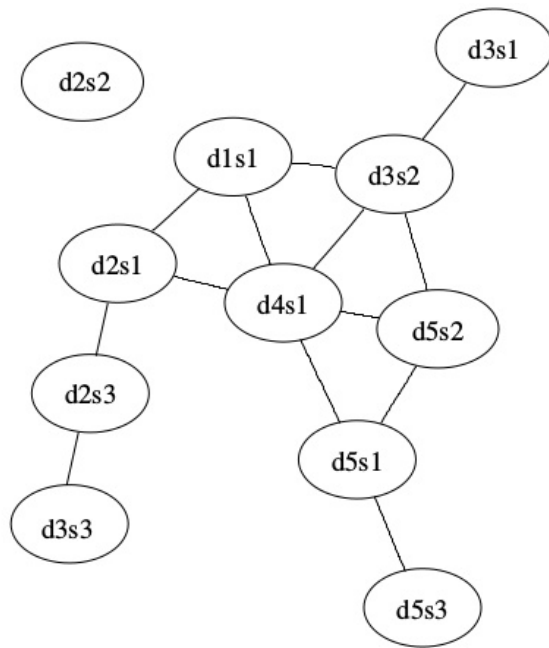
- Radev et al. 2004
  - Pick document closer to TF.IDF centroid
- Erkan et al. 2004 (“LexRank”)
  - Graph-based approach

SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

# Graph of all pair-wise cosine similarities



# Thresholded graph



SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

# Which one is more central?

## Case A:

*(A1) Syria: US, Russia get more involved in conflict*

*(A2) Syria civil war morphs into proxy war between US, Russia*

*(A3) US made weapons are turning Syria conflict into proxy war with Russia*

## Case B:

*(B1) US drops ammunition to anti-IS rebels as Syria strategy changes*

*(B2) US air drops ammunition to Syria rebels*

*(B3) US air drops ammunition to Syria rebels after strategy shift*

# User-centric evaluation

- Based on presenting the summaries to people
- Different tasks:
  - How understandable is the summary? 1..5
  - Do this summary represent these documents? 1..5
  - Given a document and a set of multi-document summaries, determine in which set this document should be
  - Etc.



# Learning from humans

*2.1.1.1. Single-document versus multi-document.* This is the *unit input parameter* or the *span parameter*, as Sparck-Jones [7] and Mani [8] respectively call it, which in simple words is the number of documents that the system has to summarize. In single-document summarization the system processes just one document at a time, whereas in multi-document summarization more than one document are processed by the system.

*2.1.1.2. Language.* Another input factor is the number of languages in which the input documents are written. So, a system can be *monolingual*, *multilingual* or *cross-lingual*. In the first case, the output language is the same as the input language. In the case of multilingual summarization systems, the output language is the same as the input language, but the system can handle a certain number of languages. In the final case of cross-lingual summarization, the system can accept a source text in a specific language and deliver the summary in another language, not necessarily the same as the input one.

*2.1.1.3. Text versus multimedia summaries.* Another important factor is the medium used to represent the content of the input document(s), as well as the output summary. Thus, we have *text*, or *multimedia* (e.g. images, speech, video apart from textual content) summarization. The most studied case is, of course, text summarization. However, there are also summarization systems that deal, for example, with the summarization of broadcast news [9] and of diagrams [10].

summaries.

*2.1.2.2. Generic versus user-oriented summaries.* This factor concerns the information a system needs to locate in order to produce a summary. Generic systems create a summary of a document or a set of documents taking into account all the information found in the documents. On the other hand, user-oriented systems try to create a summary of the information found in the document(s) which is relevant to a user query. In a sense, we can say that the query-oriented summarization systems are user-focused, adapting each time to the verbally expressed needs of the users, as viewed through the query they make or through their model (personalized summaries).

*2.1.2.3. General purpose versus domain-specific.* General-purpose systems can be easily ported to a different domain (e.g. financial, medical). This can be done, for instance, by changing the resources that characterize the domain (e.g. keywords, a domain-specific ontology), or by tuning specific parameters which concern the selection of the most appropriate techniques for the domain. On the other hand, domain-specific systems are able to process documents belonging to a specific domain.

## 2.1.3. Output factors

These factors are related to the criteria that are used to judge the quality of the resulting summary as well as with the type of summary in terms of whether this is an extract from the original document(s) or an abstraction.

# Automatic evaluation of text summarization

- Usually done with respect to a reference summary
- Problem: different people summarize things differently
  - *A Russian missile downed flight MH17*
  - *Flight MH17 crashed because of a Russian missile*
  - *MH17 crash was caused by a Russian missile*
- All of these summaries are correct!

# Bigram-based evaluation

- Reference  $\mathcal{R}$  :
  - *A Russian missile downed flight MH17*
  - *Flight MH17 crashed because of a Russian missile*
  - *MH17 crash was caused by a Russian missile*
- Sentence  $S$  :
  - *Russian missile hit flight MH17 before crash*

$$\text{ROUGE-2}(S) = \frac{\sum_{R \in \mathcal{R}} |\{\text{bigram}(b) : b \in R \wedge b \in S\}|}{\sum_{R \in \mathcal{R}} |\{\text{bigram}(b) : b \in R\}|}$$

# Bigram-based evaluation

- Reference  $\mathcal{R}$  :
  - *A Russian missile downed flight MH17*
  - *Flight MH17 crashed because of a Russian missile*
  - *MH17 crash was caused by a Russian missile*
- Sentence  $S$  :
  - *Russian missile hit flight MH17 before crash*

$$\text{ROUGE-2}(S) = \frac{\sum_{R \in \mathcal{R}} |\{\text{bigram}(b) : b \in R \wedge b \in S\}|}{\sum_{R \in \mathcal{R}} |\{\text{bigram}(b) : b \in R\}|}$$

*Exercise: compute ROUGE-2(S)*

# Indicative summaries

# Indicative summaries

State of the Union Address, 2002 vs. 2011

act afghanistan allies  
**american** attack best budget  
camps children citizens coalition  
congress continue corps country create  
danger depend destruction develop economy encourage  
enemies evil extend fight free **freedom**  
government health help history home homeland  
hope increase islamic **jobs** join lives mass  
military moment months **nation** opportunity  
peace **people** police power protect rebuild  
regimes resolve retirement **security**  
spending **states** tax **terror**  
**terrorists** thank thousands  
together tonight **training** true united  
**war** ways **weapons** women  
**work** workers **world**

President Bush, January 29, 2002

afghan ago already **american** behind  
believe best better building **business**  
care century challenge chance change child children clean  
college company compete congress country  
create cuts deficit democrats different don done  
dream economy education energy family  
future generation give goal  
**government** health help home idea  
innovation internet invest **jobs** laughter law  
life live money **nation** passed  
**people** percent possible projects race reform  
republicans research responsibility schools  
spending states step students success  
support sure tax teachers technology things together  
tonight troops willing win **work** workers  
**world** years

President Obama, January 25, 2011

Tag cloud based on absolute frequency.

# Indicative summaries

State of the Union Address, 2002 vs. 2011



President Bush, January 29, 2002



President Obama, January 25, 2011

Many words in common.

# Indicative summaries

State of the Union Address, 2002 vs. 2011

act afghanistan allies  
**american** attack best budget  
camps children citizens coalition  
congress continue corps country create  
danger depend destruction develop economy encourage  
enemies evil extend fight free **freedom**  
government health help history home homeland  
hope increase islamic **jobs** join lives mass  
military moment months **nation** opportunity  
peace **people** police power protect rebuild  
regimes resolve retirement **security**  
spending **states** tax **terror**  
**terrorists** thank thousands  
together tonight **training** true united  
**war** ways **weapons** women  
**work** workers **world**

President Bush, January 29, 2002

afghan ago already **american** behind  
believe best better building **business**  
care century challenge chance change child children clean  
college company compete congress country  
create cuts deficit democrats different don done  
dream economy education energy family  
future generation give goal  
**government** health help home idea  
innovation internet invest **jobs** laughter law  
life live money **nation** passed  
**people** percent possible projects race reform  
republicans research responsibility schools  
spending states step students success  
support sure tax teachers technology things together  
tonight troops willing win **work** workers  
**world** years

President Obama, January 25, 2011

Sometimes we care about relative frequencies more than absolute ones.

A simple method:

1. Determine  $p(\text{word}, \text{collection})$  for each word
2. Highlight words having  $p(\text{word}, \text{doc}) \gg p(\text{word}, \text{collection})$



# Colors indicate authorship

## Red = Israeli; Blue = Palestinian

fence terrorism disengagement terrorist jordan leader case bush jews past appears leaders unilateral jewish forces  
status iraq arafats line egypt green term arafat level approach abu settlers months left territory good arabs idea  
large syria suicide war strategic arab back democratic year sharons effect settlements decision bank west  
agreement majority water present mazen gaza pa sharon minister prime withdrawal israels return  
state israel process american oslo violence support security ariel peace conflict  
issue president current israeli sides palestinian israelis solution future middle  
jerusalem settlement world force plan long make issues time leadership public refugees  
east political administration pressure palestinians camp strip palestine ceasefire roadmap  
national policy government final order situation military economic hamas elections part states  
international end community territories negotiations based agreements real side united recent  
work 1967 party made movement important control authority dont hand violent borders continue change  
including clear relations problem society resolution parties building people al means move power role  
refugee ongoing intifada nations major civilians fact occupation areas talks council land struggle efforts  
hope position compromise rights stop difficult put historic opinion positions give accept reason inside law internal  
occupied americans years significant result ending things wall resistance

# Update summarization

# Update summarization



**LaSanya Rucker** @sanacardi

Russian missile-maker says its own MH17 crash investigation contradicts Dutch report - @AP  
[breakingnews.com/t/S9Q](http://breakingnews.com/t/S9Q)

🐦 a minute ago ↩ Reply ↻ Retweet ☆ Favorite



**Tynnille Kissoon** @tynnillek

RT @guardian: MH17: plane partially reconstructed as crash report blames Buk missile strike [trib.al/StEtE5j](http://trib.al/StEtE5j)

🐦 a minute ago ↩ Reply ↻ Retweet ☆ Favorite



**Obi-Wan Kenobi** @obiwankenobl

#News Flight MH17 shot down in Ukraine by Russian-built missile, report concludes [bit.ly/1LldZPm](http://bit.ly/1LldZPm) Vía @Reuters

🐦 a minute ago ↩ Reply ↻ Retweet ☆ Favorite



**newsworld** @newsynewsworld

MH17: plane partially reconstructed as crash report blames Buk missile strike via /r/worldnews [ift.tt/1WZnuBZ](http://ift.tt/1WZnuBZ)

🐦 a minute ago ↩ Reply ↻ Retweet ☆ Favorite



**ruben d vasquez** @rubendvasquez1

Dutch Safety Board: Missile exploded less than a meter (yard) from MH17 cockpit:  
[bigstory.ap.org/37c0fc99dbc74b...](http://bigstory.ap.org/37c0fc99dbc74b...) (from @AP)

🐦 4 minutes ago ↩ Reply ↻ Retweet ☆ Favorite 💬 1 more



**T██████Y** @alprazolamnufc

RT @SkyNews: Dutch investigators report finds Malaysia Airlines flight MH17 was shot down by BUK missile  
[trib.al/RnJABYP](http://trib.al/RnJABYP) <http://t.co...>

🐦 a minute ago ↩ Reply ↻ Retweet ☆ Favorite



**Malaysia - OA** @malaysia\_oa

#Malaysia Malaysia Airlines flight MH17 shot down by Buk missile, investigation finds... [bit.ly/1Lleg4Q](http://bit.ly/1Lleg4Q) -  
[ohalright.com](http://ohalright.com)

🐦 a minute ago ↩ Reply ↻ Retweet ☆ Favorite



**Punto** @libertad717

RT @AP: BREAKING: Dutch Safety Board says Buk missile fired from surface to air system downed MH17.

🐦 a minute ago ↩ Reply ↻ Retweet ☆ Favorite

# Update summarization

- Answer to the question of “what's new?”
- Input
  - Stream of documents
- Output
  - A sub-stream of documents
- Useful for mining news or social media streams
- *What is a good update?*

# What's a good update?

- Novel
  - Different from previous updates
- Timely
- Brief
- Important
  - Prevalent sentence
    - Appears in many documents
  - Central sentence
    - Similar to many other sentences

# What's an important update?

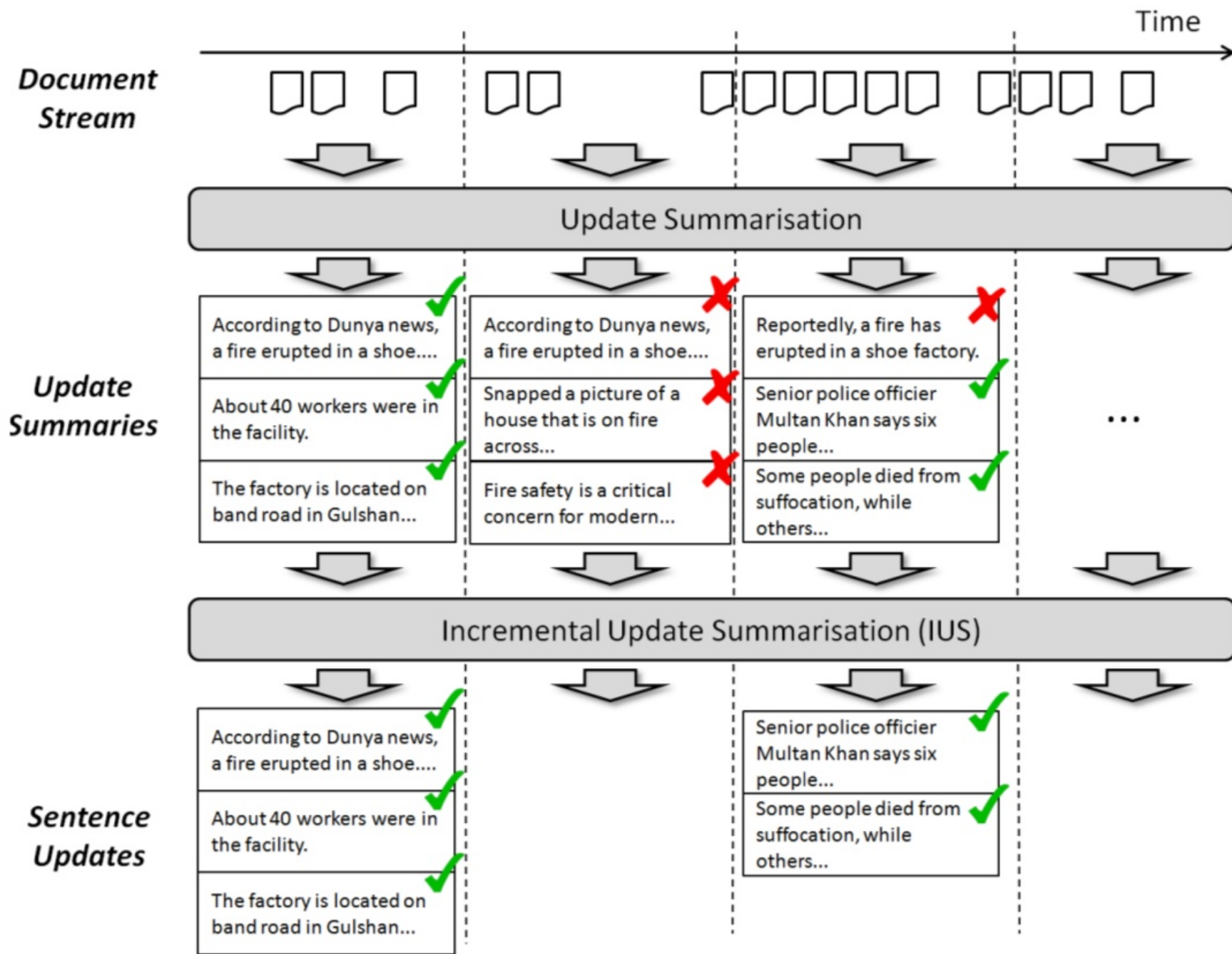
- Domain-specific approaches based on manually labeled data
  - E.g. contains keywords such as “dead”, “injured”, etc. in the domain of disasters

*Can you imagine one way in which these words can be learnt from document vectors?*

# A different approach: prevalence

		Prevalence	
		Low	High
Novelty	Low	The event is not in the news and no new information is available	The event is still in the news, but no new information is available
	High	The event is not being discussed	Important new information about the event is available

Rank	Sentence	Prevalence	Novelty
1	<i>The fire is being described as the deadliest industrial accident in Pakistan's 65-year history, and highlighted the woeful safety conditions that exist at many factories around the country.</i>	High	High
2	<i>Buildings regularly lack fire exits and basic safety equipment like alarms and sprinklers.</i>	High	High
3	<i>Fire safety induction for new staff running at 10:30am.</i>	Medium	High
4	<i>Such safety issues are common through Pakistan, where buildings also lack emergency equipment</i>	Medium	Medium





# Evaluation

- *How would you evaluate an update summarization system?*

# Evaluation

- Timeliness, coverage, succinctness
- Timestamped “nuggets” of information
  - Output updates are of the form <time, text>
  - The text of updates is compared to reference
  - A discount is applied for being late

It swings back and forth , “Thump and smash and whack.

A packed train slammed into the end of the line in Buenos Aires ' busy Once station Wednesday, injuring over 300 morning commuters, Argentina 's transportation secretary said.

Photo : Leonardo Zavattaro , Telam / AP Injured passengers from a commuter train wait to be carried away... Paramedics carry away a wounded passenger from a commuter train after a collision in Buenos Aires , Argentina , Wednesday Feb. 22, 2012.

Nugget	Dependencies	Importance
5. February 22, 2012		Low
6. Dozens killed		High
7. 550 injured		High
8. about 1,000 passengers on board the train		Low
9. the train crashed at the buffer stop		Med
10. crashed at speed of 26		Med