

“Crowdsourcing for (almost) Real-time Question Answering”

Denis Savenkov

Emory University
dsavenk@emory.edu

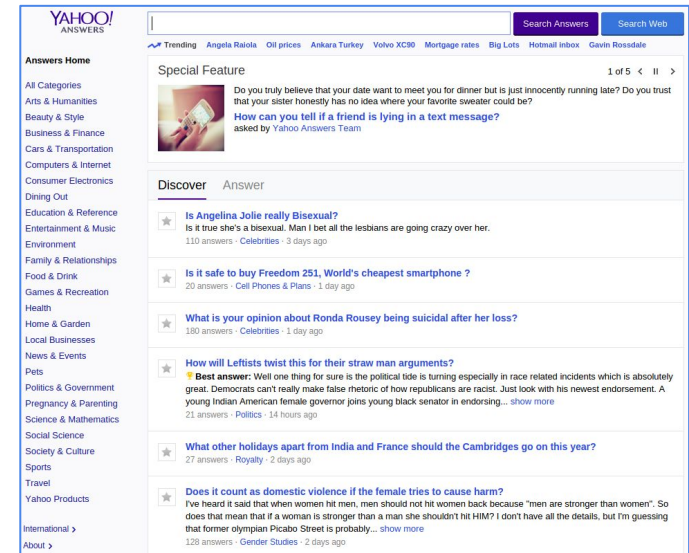
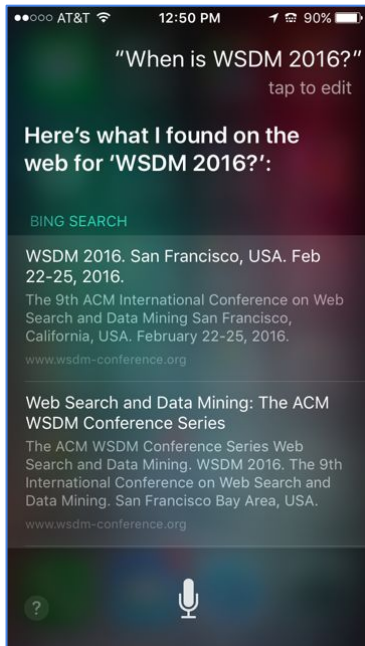
Scott Weitzner

Emory University
sweitzn@emory.edu

Eugene Agichtein

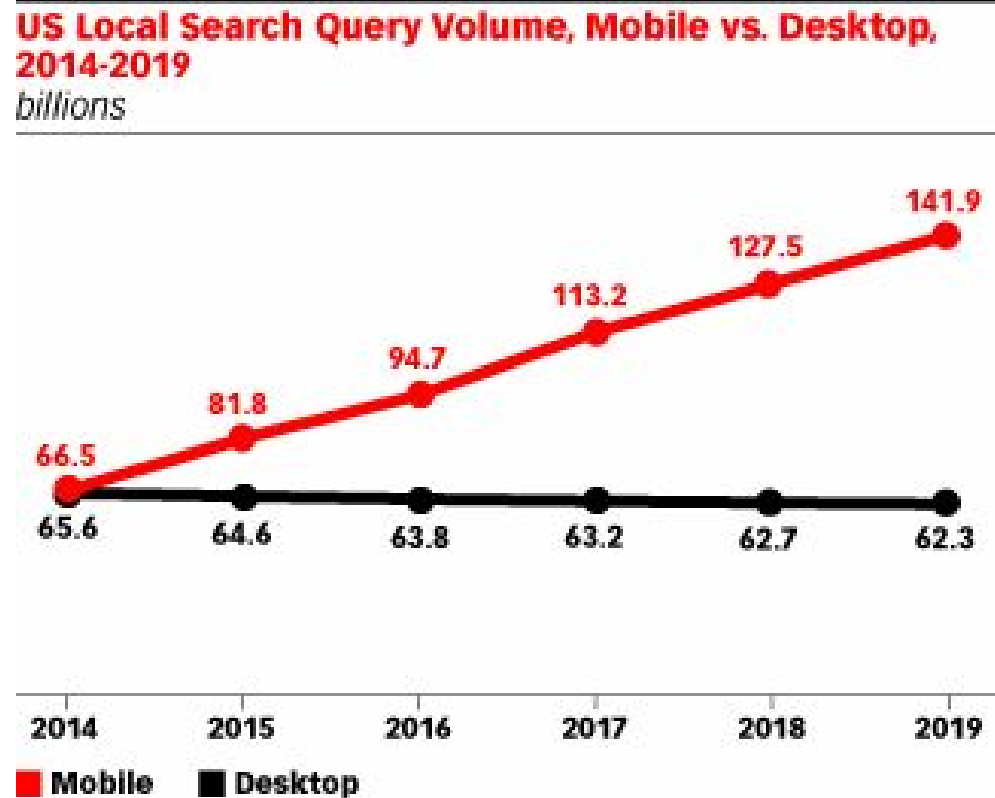
Emory University
eugene@mathcs.emory.edu

Percentage of question search queries is growing^[1]



[1] “Questions vs. Queries in Informational Search Tasks”, Ryen W. White et al, WWW 2015

And more and more of this searches are happening on mobile



Source: BIA/Kelsey as cited in company blog, May 14, 2015

190055

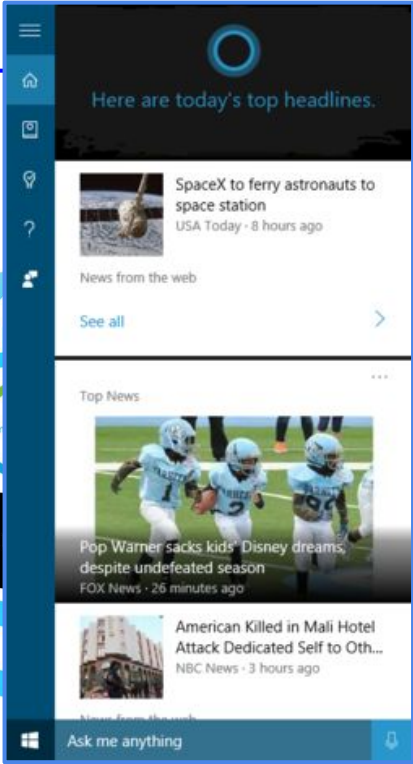
www.eMarketer.com

Intelligent Assistants today

amazon echo
Always ready, connected, and fast. **Just ask.**



Amazon Echo smart speaker with various app icons: Pandora, Spotify, Amazon, Tunein, iHeart Radio, NPR News, Calendar, Weather, Wemo, Philips Hue, Domino's, Audible.



Here are today's top headlines.

SpaceX to ferry astronauts to space station
USA Today · 8 hours ago

News from the web
See all

Top News

Pop Warner sacks kids' Disney dreams, despite undefeated season
FOX News · 26 minutes ago

American Killed in Mali Hotel Attack Dedicated Self to Oth...
NBC News · 3 hours ago

Ask me anything



Automatic Question Answering works relatively well for simple factoid questions

Google how to play go?

About 1,410,000,000 results (0.79 seconds)

The rules

1. A game of Go starts with an empty board. ...
2. Players take turns, placing one of their stones on a vacant point at each turn, with Black playing first.
3. Diagram 1 shows the position at the end of a game on a 9 by 9 board, during which Black captured a white stone at a.


More items...

[How to Play | British Go Association](http://www.britgo.org/intro/intro2.html)
www.britgo.org/intro/intro2.html

[How to Play | British Go Association](http://www.britgo.org)
www.britgo.org > An introduction to Go
The rules: A game of Go starts with an empty board. Players take their stones on a vacant point at each turn, with Black playing first position at the end of a game on a 9 by 9 board, during which Black stone at a.

[How To Play Go - Introduction](http://www.pandanet.co.jp/English/learning_go/learning_go_1.htm)
www.pandanet.co.jp/English/learning_go/learning_go_1.htm
How to play Go - Introduction to the basic rules of Go.

who was the second person to walk on the moon?



Buzz Aldrin

Engineer

Buzz Aldrin is an American engineer and former second person to walk on the Moon. He was the Apollo 11, the first manned lunar landing in history on 03:15:16 on July 21, 1969, following Michael Armstrong. He is a former U.S. Air Force officer.

[Wikipedia](#) [Twitter](#) [Facebook](#) [LinkedIn](#)

Born: Jan 20, 1930 (age 86) · Glen Ridge, NJ

Height: 5' 10" (1.78 m)

Net worth: \$10 million USD (2016)

Spouse: Lois Aldrin (1988 - 2012) · Beverly Van Jean Ann Archer (1954 - 1972)

Space missions: Apollo 11 · Gemini 12

Space agency: NASA

WolframAlpha computational knowledge engine

Is it going to rain tomorrow?

Current weather summary for Atlanta, Georgia

Today	Tomorrow	Fri	Sat
overcast wind: E at 7mph humidity: 64% 64°F	64°F 51°F	58°F 47°F	51°F 44°F

Input interpretation: precipitation forecast tomorrow

Result: rain (Thursday, April 14, 2016)

Current forecast:

Wed	Thu	Fri	Sat	Sun
rain	rain	rain	rain	rain
Apr 13	Apr 14	Apr 15	Apr 16	Apr 17

rain: 38.3% (2.8 days)

Precipitation rate (in/h):

Wed	Thu	Fri	Sat	Sun
0.08	0.04	0.04	0.04	0.04
Apr 13	Apr 14	Apr 15	Apr 16	Apr 17



(AP Photo/Jeopardy Productions, Inc.)

But for many more complex questions we still have to dig into the “10 blue links”


Google who is the father of the current US president?

All News Images Videos Shopping More Search tools

About 175,000,000 results (0.83 seconds)

Barack Obama / Father

Barack Obama, Sr.



Barack Hussein Obama, Sr., was a Kenyan senior governmental economist and the father of U.S. President Barack Obama. He is a central figure of his son's memoir, *Dreams from My Father*. [Wikipedia](#)

More about Barack Obama, Sr. Feedback

Barack Obama, Sr. - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Barack_Obama,_Sr. - Wikipedia -
Barack Obama, Sr. Barack Hussein Obama, Sr., (/ˈbærək huːˈseɪn oʊˈbɑːmɑː/; 18 June 1936 – 24 November 1982) was a Kenyan senior governmental economist and the father of U.S. President Barack Obama.

VS.

Google Why do we call Donald Trump by his last name and Hillary Clinton by her fi

All News Videos Images Maps More Search tools

About 4,670,000 results (1.23 seconds)

Why do we call Donald Trump by his last name and Hillary Clinton ...
<https://answers.yahoo.com/question/index?qid=20160608095016AADTQ7g> ▾
2 days ago - I just caught myself doing it and now I wonder why that's the trend with men vs. women in politics? We never called Obama "Barack" or Romney ...

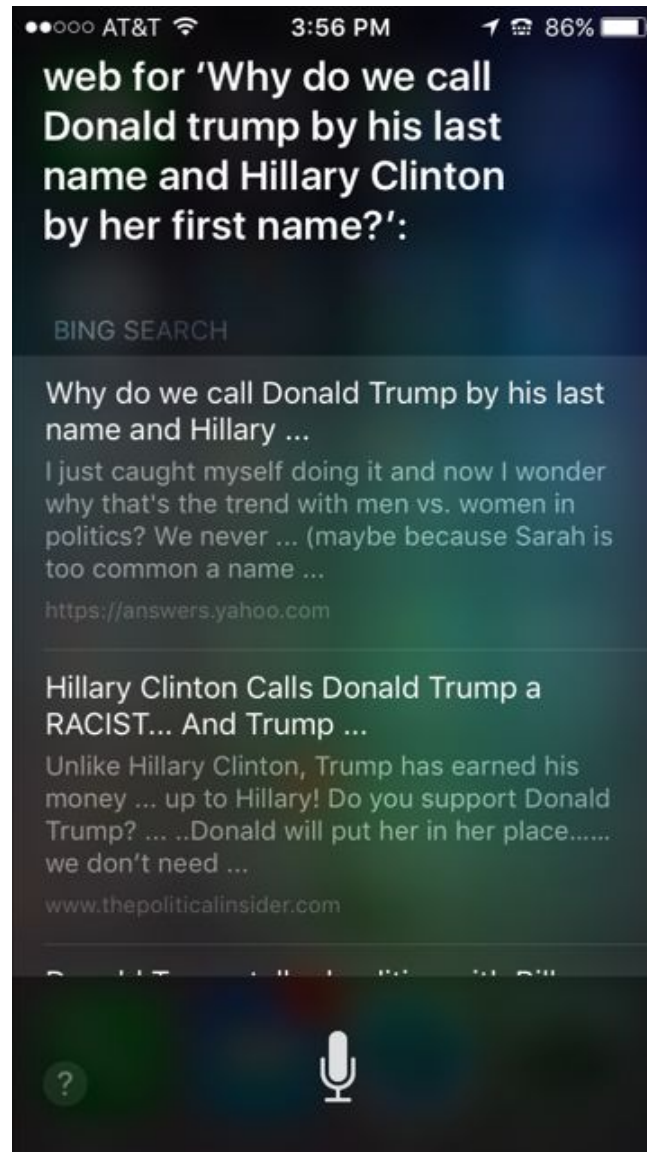
Why do we call Donald Trump by his last name and Hillary Clinton by ...
www.answerlib.org/qv/20160608095016AADTQ7g.html ▾
2 days ago - Answers of Question Why do we call Donald Trump by his last name and Hillary Clinton by her first name?:I just caught myself doing it and now ...

2016 Election: Why do people call Democrats by given names (Hillary ...
<https://www.quora.com/2016-Election-Why-do-people-call-Democrats-by-given-...> Quora ▾
Feb 9, 2016 - On the GOP side they often refer to each other by their first name. ... Jeb is called by his first name, for the same reason Hillary is: to ... is to call someone by full name on first reference, then by surname ... People who like her call her Hillary. Hillary Clinton, Bernie Sanders, Donald Trump, Ted Cruz, Ben ...

Presidential first-name trend gets Trump'd | TheHill
thehill.com/blogs/in-the-.../273333-presidential-first-name-trend-gets-trumpd ▾ The Hill ▾
Mar 17, 2016 - "It would be political malpractice for him to call himself 'Donald' or 'Don. ... store offers countless knickknacks featuring her first name, such as "Hillary for ... For Trump, it's probably better that he use his last name. ... For Clinton, "what she needed to do was differentiate herself from her husband," he explains.

What's in a name? 'Hillary' by any other name would still be ...
www.mcclatchydc.com/news/.../article24782737.html ▾ McClatchy Washington Bureau ▾
Apr 7, 2015 - "First name, last name; as long as she's getting recognized." ... Some argue that Clinton should be called by her first name because she, herself, ... a familiar frame and that is why they often promote them by first names," Kiki McLean, a senior ... Elizabeth Warren endorses Hillary Clinton, hits Donald Trump ...

... which is much harder to do on mobile



TREC LiveQA Shared Task

→ Goal:

- Answer real user questions in real-time
 - ✓ sampled Yahoo! Answers stream of questions
 - ✓ systems need to return the answer in under 1 minute
 - ✓ answers should be up to 1000 characters long
 - ✓ no restrictions on data sources

→ Last year results

No.	Run	#Answered questions	avgScore(0-3)	<i>succ@2+</i>	<i>succ@3+</i>	<i>succ@4+</i>
1	CMUOAQA	1064	1.081	0.532	0.359	0.190
2	ecnucs	994	0.677	0.367	0.224	0.086
3	NUDTMDP1	1041	0.670	0.353	0.210	0.107
4	monash-system2	1074	0.666	0.364	0.220	0.082
5	Yahoo-Exp1*	647	0.626	0.320	0.211	0.095
6	CLIP1	1079	0.615	0.326	0.204	0.086
7	emory-Out-of-mEmory	884	0.608	0.332	0.190	0.086
8	NUDTMDP3	1035	0.602	0.319	0.186	0.097
9	ECNU_ICA_2	1057	0.569	0.289	0.191	0.089
10	HIT_SCIR_QA_Grp	1086	0.522	0.291	0.168	0.063
11	ADAPT.DCU-system7	1087	0.444	0.290	0.121	0.034
12	RMIT1	1078	0.435	0.267	0.130	0.039
13	RMIT3	1082	0.415	0.251	0.126	0.038
14	NUDTMDP2	1025	0.391	0.228	0.120	0.043
15	RMIT2	1086	0.381	0.232	0.115	0.034
16	uwaterlooclarke-system4	1001	0.380	0.241	0.108	0.031
17	QU1	1082	0.256	0.163	0.070	0.023
18	DFKI-dfkiqa	1058	0.211	0.152	0.049	0.010
19	CLIP3	805	0.144	0.102	0.034	0.008
20	CLIP2	1066	0.092	0.065	0.019	0.007
21	SCU	809	0.023	0.014	0.006	0.003
Avg.		1007	0.467	0.262	0.146	0.060

Crowdsourcing for real-time QA

- Idea:
 - use crowdsourcing to help a QA system answer user questions
- Types of feedback studied:
 - Worker generated answers
 - Ratings for answer candidates
- First we want to see if we can get reasonable data from crowd workers under time pressure

Research Questions

- **RQ1.** Can crowdsourcing be used to judge the quality of answers to non-factoid questions under a time limit?
- **RQ2.** Is it possible to use crowdsourcing to collect answers to real user questions under a time limit?
- **RQ3.** How does the quality of crowdsourced answers to non-factoid questions compare to original CQA answers, and to automatic answers from TREC LiveQA systems?

Methodology: answer validation

100 questions from TREC LiveQA'15

3 answer from top-10 systems,
labelled by NIST assessors

Task: rate the quality of answers

With 1 minute time limit

- 3 workers per question
- \$0.05 per task

Without time limit

- 3 workers per question
- \$0.05 per task

LiveQA Answer Quality Scale

- **1: Bad** - contains no useful information
- **2: Fair** - marginally useful information
- **3: Good** - partially answers the question
- **4: Excellent** - fully answers the question

Validation interface

Instructions:

1. Read the given question
2. Read each of the answers and assess its quality from 1 (bad) - 4 (excellent)
3. Select one or more (if equal quality) best answers to the given question

It is possible to receive a question that is in poor taste or a question that does not make sense.

CLICK HERE WHEN YOU ARE READY TO SEE THE QUESTION

SUBMIT

Validation interface

Instructions:

1. Read the given question
2. Read each of the answers and assess its quality from 1 (bad) - 4 (excellent)
3. Select one or more (if equal quality) best answers to the given question

It is possible to receive a question that is in poor taste or a question that does not make sense.

History

What is the historical context in which the laws of Hammurabi were written?

CLICK HERE WHEN YOU ARE DONE READING

TIME LEFT: 58 SEC

SUBMIT

Validation interface

2. Read each of the answers and assess its quality from 1 (bad) - 4 (excellent)
3. Select one or more (if equal quality) best answers to the given question

It is possible to receive a question that is in poor taste or a question that does not make sense.

History

What is the historical context in which the laws of Hammurabi were written?

TIME LEFT: 40 SEC

Hammurabi is best known for the promulgation of a new code of Babylonian law: the Code of Hammurabi. This Law was written before the Mosaic Code and was one of the first written laws in the world. The Code of Hammurabi was written on a stele, a large stone monument, and placed in a public place so that all could see it...

- 1: Bad - contains no useful information
- 2: Fair - marginally useful information
- 3: Good - partially answers the question
- 4: Excellent - fully answers the question

This is the best answer

The Sunday liquor law seems to violate the spirit of the First Amendment, at least, but, because it doesn't directly address religion, would be a tough one to challenge on those grounds. As to the Ten Commandments, the courts have drawn a careful line. Displays in the context of legal history ...

- 1: Bad - contains no useful information
- 2: Fair - marginally useful information
- 3: Good - partially answers the question
- 4: Excellent - fully answers the question

This is the best answer

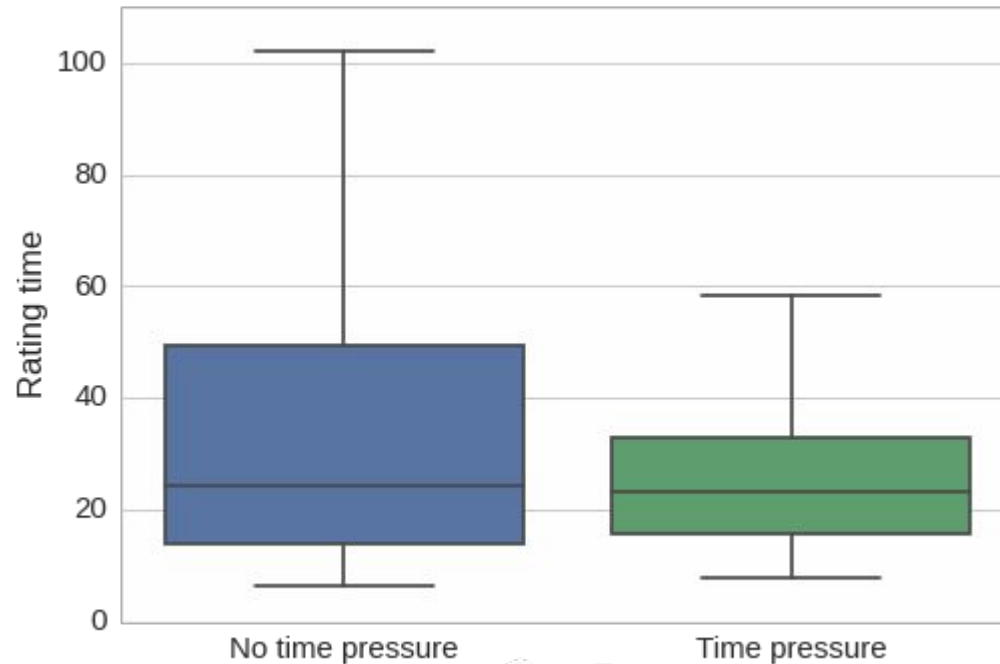
The Code of Hammurabi (also Hammurabi), the most complete and perfect extant collection of Babylonian laws, was developed during the reign of Hammurabi (r. 1792-1750 B.C.) of the first dynasty of Babylon. The code consists of Hammurabi's legal decisions, which were collected toward the end of his reign and inscribed...

- 1: Bad - contains no useful information
- 2: Fair - marginally useful information
- 3: Good - partially answers the question
- 4: Excellent - fully answers the question

This is the best answer

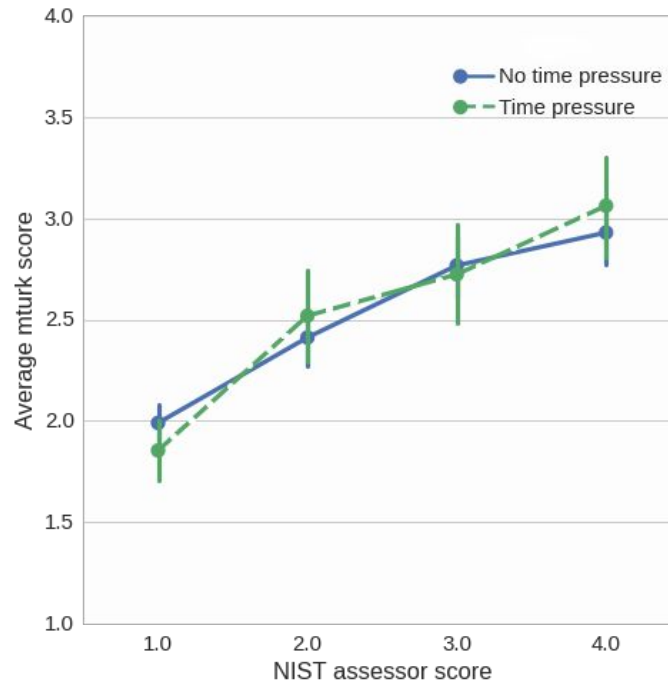
SUBMIT

Average time to judge 3 answers was 23 seconds



- ✓ Median time to judge 3 answers for a question is ~23 sec
- ✓ Time pressure forces workers to rate answers faster

Crowdsourced labels correlate well with NIST assessor scores ($\rho=0.52$)



- ✓ Workers prefer to give intermediate scores (2, 3), while NIST assessors gave more extreme scores (1 and 4)
- ✓ There is no significant difference in quality between groups with and without time pressure

Answer ratings summary

- Crowd workers can be used to obtain reliable ratings for answer candidates
- Even one minute seems to be enough to judge the quality of 3 answers to CQA questions
 - no significant rating quality loss compared to no-time pressure experiment

Methodology: answer collection

1087 questions from TREC LiveQA'15

Task: answer the given question

With 1 minute time limit

- 3 workers per question
 - \$0.10 per task
- could use web search

Without time limit

- 3 workers per question
 - \$0.10 per task
- could use web search

- Then we collected 3 ratings for each answer using crowdsourcing without time pressure
- The final quality score for the answer is the average of 3 scores

Answer Collection Interface

Instructions:

1. You will be given a question generated from a real person on the internet
2. You will have 5 minutes to answer each question
3. If you don't know the answer yourself you are allowed to browse the internet
4. If you found the answer on the internet you must provide the source (otherwise write N/A for source)
5. Use this specific link below to search for an answer, DO NOT OPEN ANOTHER SEARCH ENGINE:

WWW.GOOGLE.COM

It is possible to receive a question that is in poor taste or a question that does not make sense. Please rate each question accordingly.

[CLICK HERE WHEN YOU ARE READY TO SEE THE QUESTION](#)

SUBMIT

Answer Collection Interface

Instructions:

1. You will be given a question generated from a real person on the internet
2. You will have 5 minutes to answer each question
3. If you don't know the answer yourself you are allowed to browse the internet
4. If you found the answer on the internet you must provide the source (otherwise write N/A for source)
5. Use this specific link below to search for an answer, DO NOT OPEN ANOTHER SEARCH ENGINE:

WWW.GOOGLE.COM

It is possible to receive a question that is in poor taste or a question that does not make sense. Please rate each question accordingly.

Question:

60

Anybody know how THE CODING INVOLVED to add the feature to tag someone on a social media website?

I want the person to get a notification. So if I used @johndoe I want johndoe to get a notification. How do I do that?

[CLICK HERE WHEN YOU ARE DONE READING](#)

SUBMIT

Answer Collection Interface

Question: TIME LEFT: 34 SEC

Anybody know how THE CODING INVOLVED to add the feature to tag someone on a social media website?

I want the person to get a notification. So if I used @johndoe I want johndoe to get a notification. How do I do that?

Does the way the question is worded make sense?

yes

no

Are you familiar with this topic?

yes

no

Write Your Answer Below:

1000 Character Limit

Answer Source:

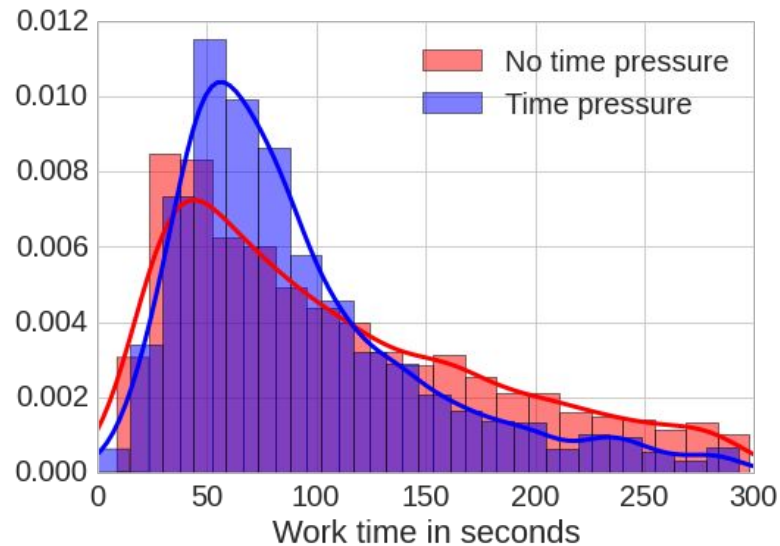
Answer Source

TIME LEFT: 34 SEC

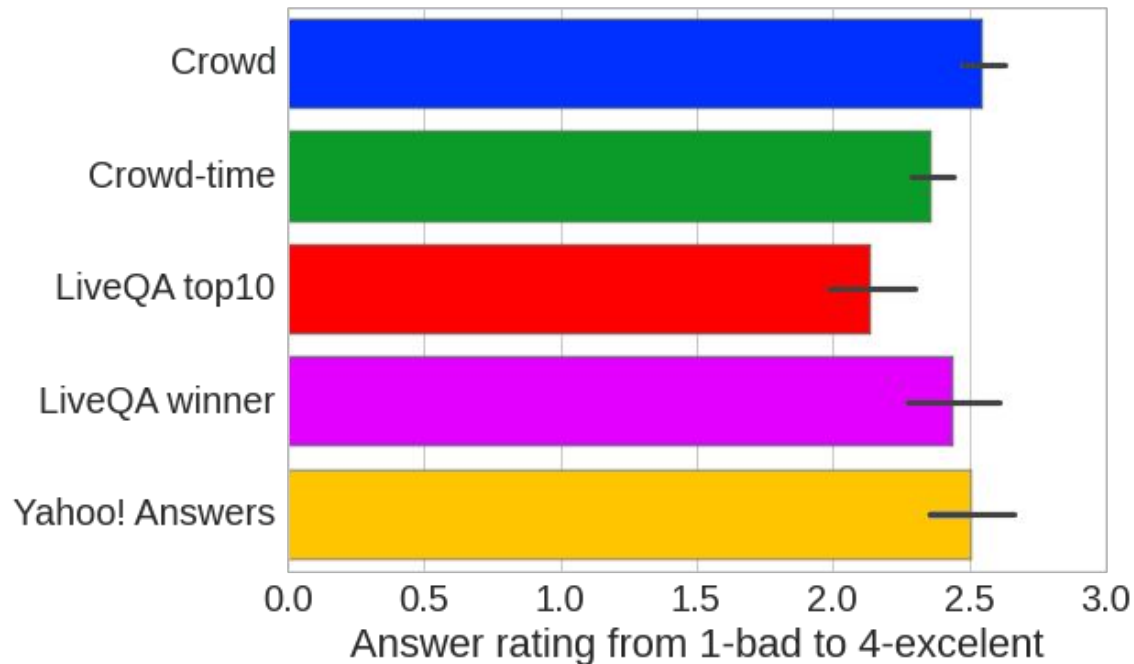
SUBMIT

Answer Crowdsourcing statistics

	Yahoo! Answers	mTurk	mTurk ≤ 1 min	LiveQA'15 winning system
% answered	78.6%	100.0%	100.0%	97.8%
Length (chars)	354.96	190.83	126.65	790.41
Length (words)	64.54	34.16	22.82	137.23

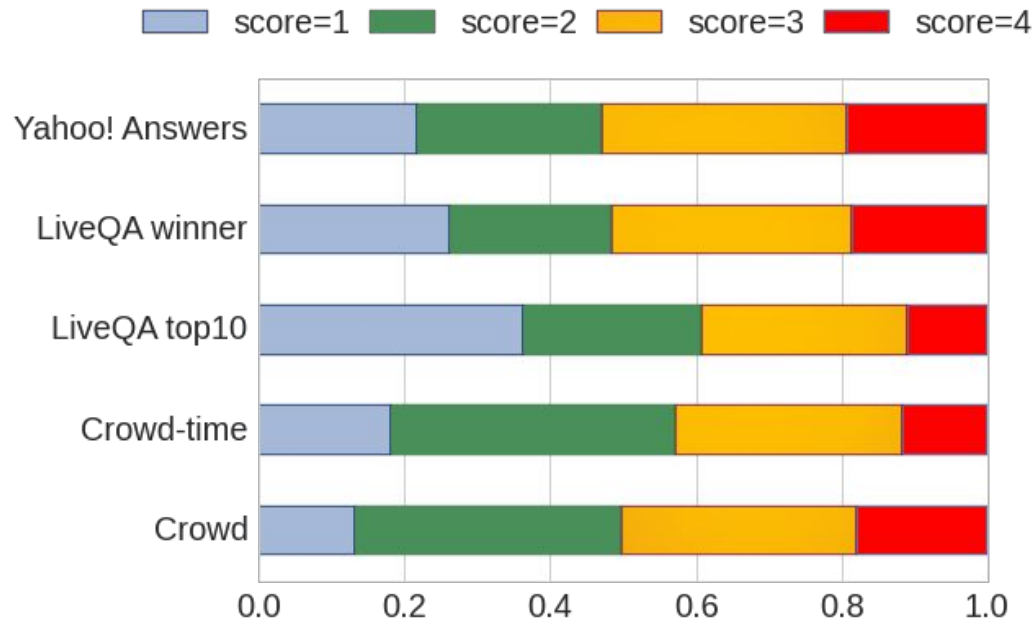


Average quality scores of answers



- ✓ Crowdsourced answers are comparable in quality to community generated responses
- ✓ Answers obtained from a crowd under 1 minute time limit are significantly better than those of a top-10 automatic QA system

Answer Crowdsourcing statistics



- ✓ Crowdsourced answers tend to be relevant, but of average quality
- ✓ Automatically generated answers are more often either not-useful or good

Example Answers

- General Opinions and Recommendations

- What is a great free(or cheap) video player/library combo that somewhat functions like Itunes?

- ✓ Media Player Classic might be what you are look...

- Questions requiring certain expertise

- Are gold cap conures good birds? I'm looking for sweet, loving ect.?

- ✓ Apparently they are adorable. Their only imperfection is that they are loud.

- ✓ Conures can be very loud, obnoxiously ear peircingly loud.

- Is Gotu Kola a good herb for mental health? How long does it take to work??

- ✓ yes

- Which LGA 775 processor can I upgrade to?

- ✓ You have to make sure what you want to upgrade to is compatible or it won't work with your machine. There should be a list on any processor that tells what it works with.

- What is OPEC46LCZ?

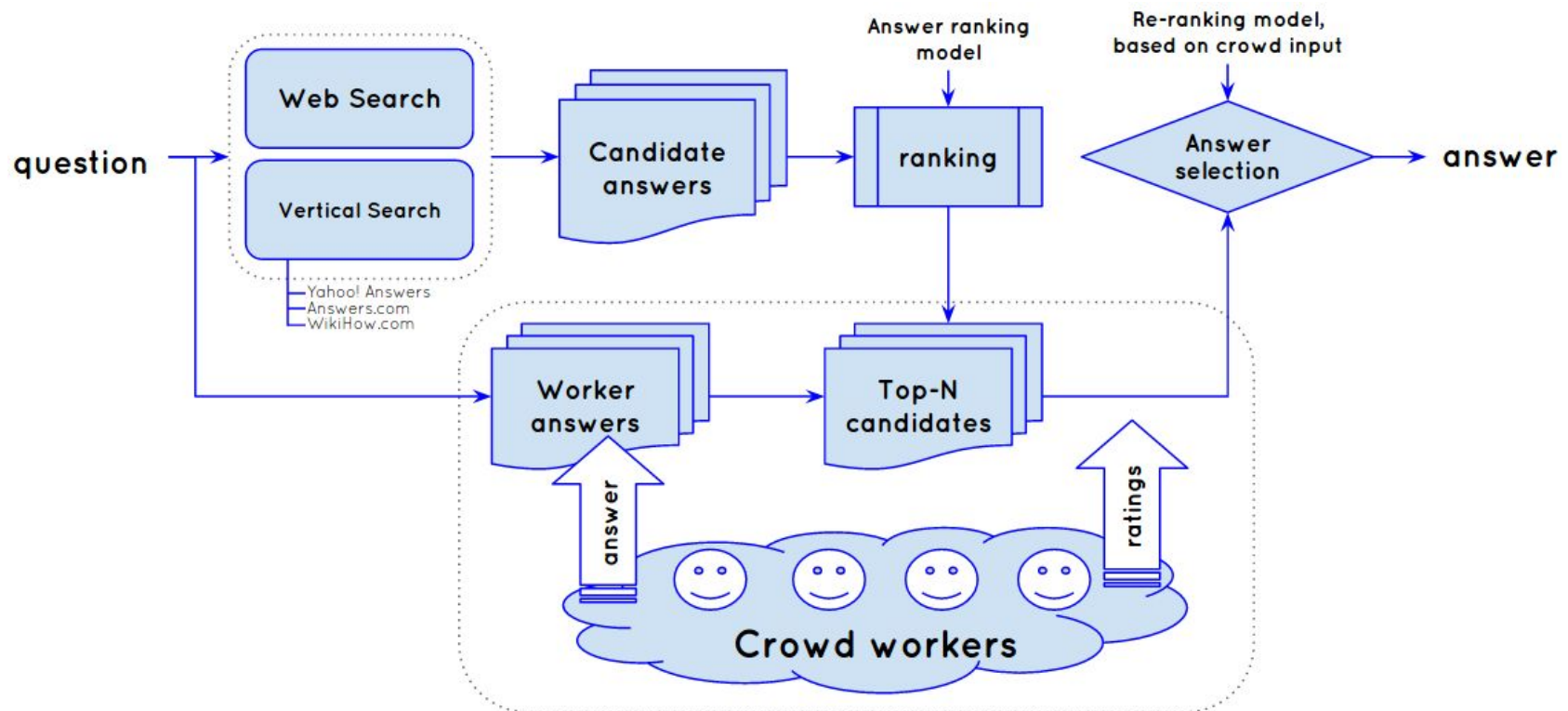
- ✓ OPEC is about oil and nothing to do with heart disease.

- ✓ There is no answer available to this question

Conclusions

1. Crowd workers are capable of reliably validating ~3 answers to a question, even with 1 minute time limit
2. Even one minute appears to be enough for crowd workers to provide a response to real user questions, such as those posted to CQA websites
3. The quality of crowdsourced answers on average was comparable to the CQA community answers, and even with time limit crowdsourcing can be useful for an automated QA system

Crowd-powered Real-time Question Answering



- ✓ We integrated crowdsourcing into a real-time automated QA system, that participated in TREC LiveQA'16
 - To rate candidate answers
 - To provide additional candidates

Crowd-powered Real-time Question Answering

INSTRUCTIONS Question time: 00:39 HIT TIME: 14:49

Politics

What is a male 'First Lady' called if there was a female President?

If you can provide a good response, please type it here...

SUBMIT >

I think it will be something gender neutral and will be used for males and females in the future. Like someone here mentioned 'first spouse' would be a good one

show all

- 1: Bad - contains no useful information
- 2: Fair - marginally useful information
- 3: Good - partially answers the question
- 4: Excellent - fully answers the question

First Lady is an unofficial title used for the wife or hostess of a non-monarchical head of state or chief executive.[1][2][3] The term is also used to describe a woman seen to be at the top of her profession or art.[4] Collectively, the President of the United States and his spouse are known as the...

show all

- 1: Bad - contains no useful information
- 2: Fair - marginally useful information
- 3: Good - partially answers the question
- 4: Excellent - fully answers the question

First Gentleman, whether he is or not. An ex-president is always called 'President So-and-so', so if Hilary Clinton is elected President, there will be two President Clintons in the White House. They will be referred to as President Bill Clinton and President Hilary Clinton according to custom, so i...

show all

- 1: Bad - contains no useful information
- 2: Fair - marginally useful information
- 3: Good - partially answers the question
- 4: Excellent - fully answers the question

© 2016 Intelligent Information Access Lab, Emory University

- ✓ Workers accept Amazon mTurk HIT, that lasts 15 minutes
- ✓ Within this time, they are waiting for questions to arrive
- ✓ When a question arrives, a crowd worker could do the following:
 - Provide his answer to the question if she has one (optional)
 - Rate answers as they appear on the screen

Average rating of answers

	Average score per question	Average precision
Automated system	2.321	2.357
Automated system + crowdsourcing (answers+ratings)	2.550 ⁺	2.556 ⁺
Ratings only	2.432	2.470
Answers only	2.459	2.463

- ✓ Crowdsourcing for real-time question answering significantly improves the performance
- ✓ Ratings and answers both contribute to the overall quality gain

- Crowdsourcing can significantly improve the performance of a near real-time question answering system by providing additional answers and rating existing answer candidates
- Future work:
 - Cost & Scalability:
 - Optimizing the number of workers per task
 - Selective crowdsourcing: e.g. using query performance prediction techniques
 - What are other useful types of feedback a crowd can provide to QA system
 - e.g Question summary
 - expected phrases in the answer

Thank you!