

Assignment 5

Fall 2017

CS834 Introduction to Information Retrieval

Dr. Michael Nelson

Orkun Krand

December 15, 2017

1 Question 10.5

1.1 Question

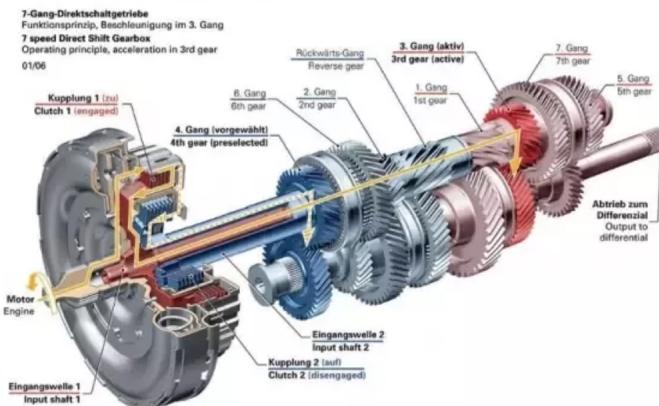
Find a community-based question answering site on the Web and ask two questions, one that is low-quality and one that is high-quality. Describe the answer quality of each question.

1.2 Answer

I decided to use [Quora](#) to answer this question. My high quality question was: "How does a clutch work?" This is a high quality question because it is asking for serious information on how a car part operates. As expected, the answers were serious as well with some providing images to support the answer. There were no jokes or funny remarks. Because Quora classifies questions into topics, nobody thought the question was about women's purses or the 1990s rock band.

The clutch is the part that basically connects the engine to the gearbox.

(Shown with gearbox here)



Basically, from the left is where the engine would be, and the right is where the shaft connects to the wheels.

The clutch, when engaged, disconnects the engine to the gearbox so the gears (gear ratio) can be changed without damaging the gears.

The clutch is operated using the most left of the 3 pedals of a manual car. If the car is automatic, it's slightly different.

Figure 1: How does a clutch work

My low quality question was: "Why do motorcycle riders always feel compelled to wave at other motorcycle riders in passing, or when coming to a stop light?" The reason the words "feel compelled" is used is to show that this question isn't really looking for a valid reason but it is more of a fun question. It's not a technical question that has one answer, it also isn't a universal case but the

United States or North America isn't mentioned in the question. The purpose of the question was to ask about the origin of the motorcyclist wave, but because of the way it was worded and the way the wave actually works (or doesn't), it can't be considered a high quality question.



Figure 2: Low quality question

The answers posted were mostly entertaining with one person listing reasons why people who own a certain brand of bike won't wave back. I can safely say that I did not learn anything from the answers to this question.

There are many great answers posted here, but in case you were wondering, here is a tongue in cheek list of reasons bikers DON'T wave, by brand of bike (I'm a BMW rider by the way - all true I tell you)

Top Ten Reasons Why Harley Riders Don't Wave Back.

10. Afraid it will invalidate warranty. 9. Leather and studs make it too heavy to raise arm. 8. Refuse to wave to anyone whose bike is already paid for. 7. Afraid to let go of handlebars because they might vibrate off. 6. Rushing wind would blow scabs off the new tattoos. 5. Angry because just took out second mortgage to pay luxury tax on new Harley. 4. Just discovered the fine print in owner's manual and realized H-D is partially owned by Honda. 3. Can't tell if other riders are waving or just reaching to cover their ears like everyone else. 2. Remembers the last time a Harley rider waved back, he impaled his hand on spiked helmet. 1. They're too tired from spending hours polishing all that chrome to lift their arms.

Top Ten Reasons Why Gold Wing Riders Don't Wave Back.

10. Wasn't sure whether other rider was waving or making an obscene gesture. 9. Afraid might get frostbite if hand is removed from heated grip. 8. Has arthritis and the past 400 miles have made it difficult to raise arm. 7. Reflection from etched windshield momentarily blinded him. 6. The espresso machine just finished. 5. Was actually asleep when other rider waved. 4. Was in a three-way conference call with stockbroker and accessories dealer. 3. Was distracted by odd shaped blip on radar screen. 2. Was simultaneously adjusting the air suspension, seat height, programmable CD player, seat temperature and satellite navigation system. 1. Couldn't find the "auto wave back" button on dashboard.

Figure 3: Reasons motorcyclists don't wave back

As expected, the higher quality question returned higher quality answers that were in depth, on point

and were written as clearly as possible to avoid any confusion to the reader. The low quality question received answers that were not serious, or not related.

2 Question 10.6

2.1 Question

Find two examples of document filtering systems on the Web. How do they build a profile for your information need? Is the system static or adaptive?

2.2 Answer

[Ebay](#) is an online shopping website that is used worldwide to purchase all sorts of goods from pencils to used car parts to baby clothes. It keeps track of what the user has been searching for and buying to create a profile of them. They use this profile to recommend more things to buy to the user as shown in the picture below. It is not as precise as I'd like but it still does a good enough job that I click the items that come up as recommended, probably once a month. It fails to distinguish brands. The bike I am looking for a fairing for is a Yamaha, yet it keeps showing me Suzuki fairings.

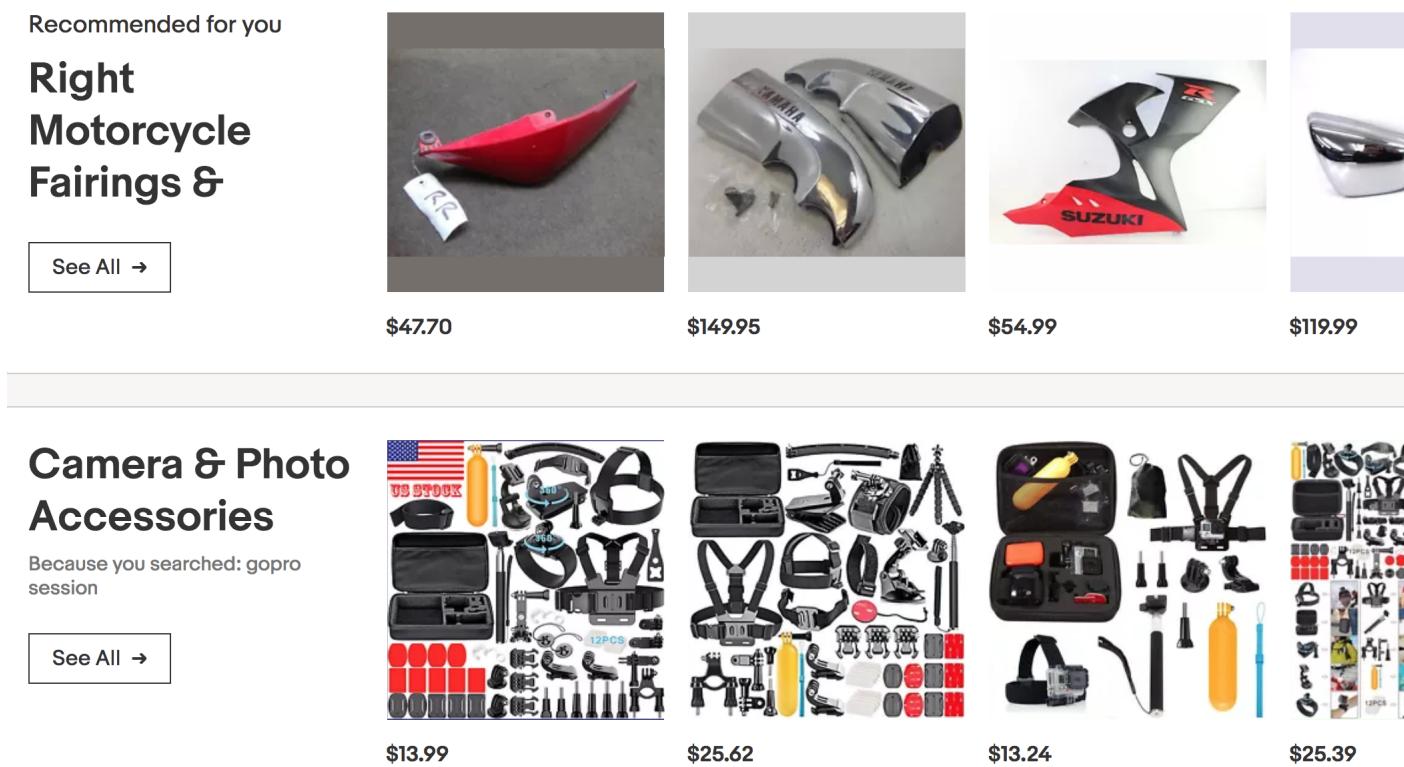


Figure 4: Ebay recommended items

Classifying new items is simplified thanks to their very intensive categorization of the items. Users

can create their ads by finding something similar for sale and simply clicking the button below the picture that says “Have one to sell? Sell now”. This will fill out some fields for you depending on which ad you navigated from. But the main way to tell Ebay what you are selling is to use the categories listed when creating a new ad. This makes sure Ebay knows exactly what you are posting so it can decide to notify or not notify people looking for that item.

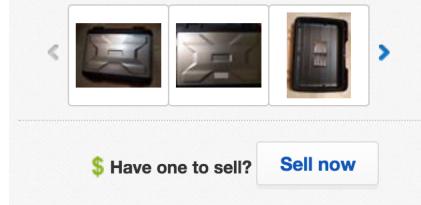


Figure 5: Ebay “have one to sell” button

*Category

eBay Motors > Parts & Accessories > Motorcycle Accessories > Luggage > Other Luggage

Suggested categories | Recently used categories | Search categories | Browse categories

- eBay Motors > Parts & Accessories > Motorcycle Accessories > Luggage > Other Luggage
- eBay Motors > Parts & Accessories > Scooter Parts > Luggage & Luggage Racks > Luggage Racks
- eBay Motors > Parts & Accessories > Motorcycle Parts > Antique, Vintage, Historic
- eBay Motors > Parts & Accessories > Motorcycle Accessories > Luggage > Saddlebags

Figure 6: Ebay category box

YouTube is a video sharing platform owned by [Google](#). YouTube is a little different from your average e-commerce websites (Amazon, Ebay etc.). There are two different ways to view YouTube. First is the every day person’s YouTube activity which is searching for a topic and watching videos about said topic. The second is following a content creator in which the user would watch videos made by the content creator even if he/she doesn’t really care about the topic of that specific video.

But the overall idea is the same in that each user will have a profile that consists of the types of videos they like to watch, the channels they have subscribed to and kind of videos they have skipped. For example it knows I like motorcycle videos. Therefore, it recommends motorcycle videos to me. It doesn’t care as much about the specific topic of the video if I’m subscribed to the channel. YouTube is also biased thanks to millions of other viewers rating videos and making certain people more popular than others. So if someone popular on YouTube publishes a new video, it would appear in my home page even if I am not subscribed to that person.

For new videos, it has an weighted evaluation function where it gives different properties of a video different weights. Since YouTube is a video sharing platform and videos aren’t easy to crawl, it goes by the title, description, and tags. For example I make motorcycle videos on YouTube in Turkish. There are two types of viewers that watch my content. One is people who speak Turkish who are interested in motorcycles and the second is people who speak Turkish who are interested in life in the United States. After a year of doing it, I’ve noticed that YouTube gives a heavy weight to the video titles. Therefore, I try to include the word “Amerika” in my video titles to reach the latter type of audience. Basically trying to play tricks on the decision mechanism to boost my channel. Below is a screenshot of my YouTube Analytics page which shows what people searched for to find my channel. This is where I got the idea.

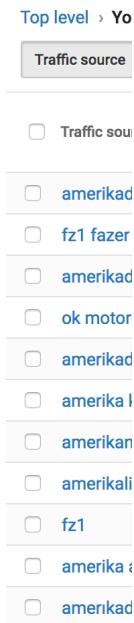


Figure 7: YouTube traffic sources search results page

3 Question 11.5

3.1 Question

How many papers dealing with term dependency can you find in the SIGIR proceedings since 2000?
List their citations.

3.2 Answer

Using [Google Scholar](#)'s advanced search tool, I was able to find all the papers that were published in SIGIR after 2000 that deal with term dependency. I used the advanced search tool because I am not very competent about using the specific symbols to tell Google what I want so I let the tool do it for me. Since the question said “dealing with” term dependency, I figured I’d include a couple variations of the search term so I searched for “term dependency” or “term dependencies” or “term dependence”. I’m aware that stemming would probably take care of this for me but I didn’t want to take chances. A screenshot of the search is below. The list of the citations can be reached at [this link](#).

4 Question 11.9

4.1 Question

Find a demonstration of a question answering system running on the Web. Using a test set of questions, identify which types of questions work and which don’t on this system. Report effectiveness using MRR or another measure.

The screenshot shows a Google Scholar search interface. The search query in the bar is "term dependencies | term dependency | term dependence source:SIGIR". Below the search bar, it says "About 223 results (0.05 sec)". On the left, there's a sidebar with filters: "Articles" selected, "Any time", "Since 2017", "Since 2016", "Since 2013", "Custom range...", date input fields (2000 to 2023), and a "Search" button. Below these are sorting options: "Sort by relevance" (selected) and "Sort by date". There's also a checkbox for "include citations". The main content area lists three research papers:

- A Markov random field model for term dependencies** by D Metzler, W.B. Croft, et al., from the 28th annual international ACM SIGIR, 2005. It discusses estimating dependencies in relevant and non-relevant documents.
- Incorporating term dependency in the DFR framework** by J Peng, C. Macdonald, B. He, V. Plachouras, et al., from the 2007 international ACM SIGIR. It talks about modeling dependencies between query terms.
- Dependence language model for information retrieval** by J. Gao, J.Y. Nie, G. Wu, G. Cao, et al., from the 27th annual international ACM SIGIR, 2004. It discusses estimating dependencies statistically on the scale of whole documents.

Each paper entry includes a star rating, citation count (e.g., "Cited by 774"), "Related articles", and "All versions".

Figure 8: Google Scholar search for SIGIR papers after 2000 dealing with term dependency

4.2 Answer

For this question I picked [MIT's START](#). It is different than a search engine in that it tries to give you one, correct answer for your question instead of taking you to a list of pages that might have the answer you are looking for. But it is also similar to a search engine in that it will collect its answer from different resources. For example just entering “Old Dominion University” in the search bar and information from Wikipedia, and U.S. News. It also takes a little while (couple seconds) to present answers for your query.

It requires and produces English text so any other language will likely fail to produce results. It also fails to answer questions that are subjective such as “What is love?”, even though I receive a definition when I query “love”. Unlike Google, it seems that it doesn’t have a concept of popular query. When I query “How does ABS work?”, I get information about the New Zealand rugby team All Blacks. I am expected to go over the other possible meanings of ABS and find the one I am looking for. So it is safe to say that abbreviations aren’t the best way to go about using this system.

So in order to calculate effectiveness, I will use the following sample queries:

- How does clutch work
- What is black eyed pea (referring to the type of beans)
- How does ABS work
- The Killers (referring to the band)
- Ford

Since these are terms that have multiple meanings, for most cases, I received a list of possible meanings that seemed to be retrieved from Wikipedia. Below are screenshots of what I received for each query. In two of the five, I received what I was looking for immediately. This has to do with me trying to

trick the system and querying things that have more than one meaning. Not many people search for black eyed pea, the bean.

==> How does clutch work

Clutch

A **clutch** is a mechanical device which engages and disengages power transmission especially from **driving shaft** to driven shaft.

In the simplest application, clutches connect and disconnect two rotating shafts (drive shafts or **line shafts**). In these devices, one shaft is typically attached to an engine or power unit (the driving member) while the other shaft (the driven member) provides output power for work. While typically the motions involved are rotary, linear motion is also possible.

I know about 11 more terms called "Clutch": [Clutch \(pin fastener\)](#), [Clutch \(mascot\)](#), [Clutch \(magazine\)](#), [The Clutch](#), [Clutch \(web series\)](#), [Clutch \(band\)](#), [Clutch \(literary character\)](#), [Clutch \(Clutch album\)](#), [Clutch \(Peter Hammill album\)](#), [Clutch \(eggs\)](#), and [Clutch \(G.I. Joe\)](#)

Source: [Wikipedia](#)

Figure 9: START search results for clutch

==> What is black eyed pea

The Black Eyed Peas

The **Black Eyed Peas** (originally simply **Black Eyed Peas**) are an American group, consisting of rappers [will.i.am](#), [apl.de.ap](#), [Taboo](#), and [Fergie](#). Originally an [alternative rock](#) group, they subsequently changed their musical sound to [pop](#) and [dance-pop](#) music.^[2] Although the group was founded in [Los Angeles](#) in 1995, it was not until the release of their third album [Elephunk](#) in 2003 that they achieved high record sales. Since that time, the group has sold an estimated 76 million records (35 million albums and 41 million digital downloads), making them one of the [world's best-selling groups of all time](#).^[3] According to [Nielsen SoundScan](#), the Black Eyed Peas were the second-best-selling artist/group in 2011, with over 42 million sales as of the end of 2011.^[4]

I know about two more terms called "Black Eyed Pea": [Black-eyed pea](#) and [Black-eyed Pea \(restaurant\)](#)

Source: [Wikipedia](#)

Figure 10: START search results for black eyed pea

Overall, the mean reciprocal rank from these searches was: $\frac{1+\frac{1}{2}+\frac{1}{31}+0+\frac{1}{5}}{5} = 0.346$

5 Question 11.11

5.1 Question

Look at a sample of images or videos that have been tagged by users and separate the tags into three groups: those you think could eventually be done automatically by image processing and object recognition, those you think would not be possible to derive by image processing, and spam. Also decide which of the tags should be most useful for queries related to those images. Summarize your findings.

5.2 Answer

I chose the popular image based social media [Instagram](#) to answer this question. Users upload their images to Instagram and add tags using the # sign. Tags are short phrases that explain the image uploaded. Many users use the tags to reach a wider audience. Instagram works on likes and follows. A picture with more likes will likely be shown to more people and follower counts are considered for ranking.

====> How does ABS work

I can answer your query, "How does ABS work", for this 46 terms called "abs":

- [Abs District](#)
- [Abs \(surname\)](#)
- [Amalgamated Broadcasting System](#)
- [Association of Broadcasting and Allied Staffs](#)
- [Cis-abienol synthase](#)
- [ABS Global](#)
- [Abs \(Yemen\)](#)
- [ABS Free Dish](#)
- [Akita Broadcasting System](#)
- [Australian Bureau of Statistics](#)
- [ABS-CBN](#)
- [Aston Business School](#)
- [Rectus abdominis muscle](#)
- [Abz Love](#)
- [Abby Sciuto](#)
- [American Bible Society](#)
- [American Bureau of Shipping](#)
- [Automatic block signaling](#)
- [Abingdon Boys School](#)
- [Absorbance](#)
- [Absolute case](#)
- [Asset-backed security](#)
- [Able seaman \(rank\)](#)
- [ABS Capital Partners](#)
- [Banu Abs](#)
- [AB'S](#)
- [Ambonese Malay](#)
- [Absolute value](#)
- [Ankaferd BloodStopper](#)
- [Anti-lock braking system](#)
- [American Boy Scouts](#)
- [Acrylonitrile butadiene styrene](#)
- [Alternative Bank Schweiz](#)
- [Archbold Biological Station](#)
- [Able seaman](#)
- [American Bladesmith Society](#)
- [Amniotic band constriction](#)
- [Animal Behavior Society](#)
- [ABS network](#)
- [ABS \(TV station\)](#)
- [Albino Blacksheep](#)
- [ABS Building Society](#)
- [ABS \(satellite operator\)](#)
- [Abu Simbel Airport](#)
- [Michael Absalom](#)
- [Danieli](#)

Figure 11: START search results for ABS

marketing purposes. That's why spam tags such as #likeforlike and #likeforfollow are very common on accounts that are trying to advance in the "Instagram world". While some tags can be generated automatically, some aren't easy to identify via image processing and other techniques. Tags such as #smile can be generated by image processing. But it would be hard to automatically generate tags that convey emotion such as #happy which returns 407 million results.

A sample of these tags can be seen in the screenshots below. Users try to use as many tags to reach as many people as possible. The second picture belongs to an account called [onherbike](#). It is the Instagram account of a motorcycle travel blogger. She uses Instagram tags to build her audience which she needs to be able to do her travels. She is sponsored by many companies including [BMW](#)

I can answer your query, "The Killers", for this 45 terms called "killer":

- Killer (philately)
- Killer (pool)
- Killer (Kiss song)
- The Killer (1953 film)
- Killer (nickname)
- Killer Kane
- Taggart
- Killer Kowalski
- Le Tueur
- The Killer (1972 film)
- The Killer (play)
- Killer (1992 film)
- Killer (Finnish band)
- Killer (video album)
- Killer (Alice Cooper album)
- Tales from the City
- The Killer (2006 film)
- Killer application
- Killer (Belgian band)
- The Killer (The Vampire Diaries)
- This Man Must Die
- Killer (drag performer)
- Killer NIC
- Ghostface (Scream)
- Daniel Killer
- Killer (Adamski song)
- Killer: A Journal of Murder
- Killer: The Game of Assassination
- The Killer (Wilson novel)
- Tobias Killer
- Burladingen
- Wink murder
- The Killer (comics)
- Killer yeast
- Death receptor 5
- Bob Konovsky
- Killer (Swiss band)
- Operation Sundevil
- The Killer (Shed album)
- Assassin (game)
- Killer: Dead or Alive
- Killer (1998 film)
- The Killer (Impious album)
- Killer (Tech N9ne album)
- Killer sudoku

Figure 12: START search results for The Killers

[Motorcycles](#) who probably require her to use these tags on her pictures. The third picture is a screenshot of many spam tags. The content of the picture is not very relevant in this case as the tags don't reflect anything in the picture. They are simply used to lure people.

This user is trying to reach an audience that is interested in motorcycles, especially cafe racers. Even though it is hard to tell much about the motorcycle from this picture, it is generally visually pleasing, so maybe more general tags such as #adventure, #reflection would have increased the range (how many people it reaches) of this post. Longer tags are usually not preferred as it kind of beats the purpose of tags so the tags that benefited this post the most are likely the shorter ones such as #motorcycle, #caferacer and #biker.

The user that posted this picture is sponsored by BMW so it is normal to see all the BMW tags in all her pictures. Location tags such as #istanbul are very popular which has likely benefited this post

==> ford

I assume that you wanted to know what Ford is.

Ford is a city located in the USA, in the state of Minnesota.

Source: U.S. Census Bureau

Ford is a city located in the USA, in the state of Wisconsin.

Source: U.S. Census Bureau

Ford is a city located in the USA, in the state of Kansas.

Source: U.S. Census Bureau

Source: START KB

Ford Airport (IMT)

Iron Mountain MI, US 49801

Latitude: 45.815556, Longitude: -88.11417

Source: FlightStats

Ford Motor Company

The Ford Motor Company (commonly referred to simply as "Ford") is an American multinational automaker headquartered in Dearborn, Michigan, a suburb of Detroit, Michigan, United States. It was founded by Henry Ford and incorporated on June 16, 1903. The company sells automobiles and commercial vehicles under the Ford brand and most luxury cars under the Lincoln brand. Ford also owns Brazilian SUV manufacturer, Troller, and Australian performance car manufacturer FPV. In the past, it has also produced tractors and agricultural components. Ford owns an 8% stake in Aston Martin of the United Kingdom, [needs update] and a 49% stake in Jiangling of China[citation needed]. It also has a number of ventures, one in China (Changan Ford), one in Taiwan (Ford Lio Ho), one in Thailand (AutoAlliance Thailand), one in Turkey (Ford Otosan), and one in Russia (Ford Sollers). Ford is listed on the New York Stock Exchange and is controlled by the Ford family, although they have minority ownership (but majority of the voting power).^{[4][3]}

I know about 48 more terms called "Ford". You don't really want me to list them all, do you?

[List them all](#)

Source: Wikipedia

Which person in The Internet Movie Database do you mean:

- Faith Ford
- Harrison Ford (II)
- Lita Ford

I know about 146 more persons called "Ford". You don't really want me to list them all, do you?

Figure 13: START search results for Ford

more than the others. #makelifeareide doesn't seem like an optimal tag as it doesn't convey much information in itself. Upon clicking on it, I noticed that this tag is mostly used by BMW motorcycle riders who already use all the other similar tags. So most of the tags that are used in this post have the same audience and therefore can be ignored. #motorcycle, just like the first example is a useful tag that will draw people. #spiritofGS, probably not that important for this post.



Figure 14: More tags = More people



Figure 15: Users use tags to build up an audience to promote their product

#forsale #sale #sold #soldout #buy #black
#helm #helmet #shoei #race #racing
#like4like #likeforlike #follow4follow
#follow4follow #like #follow #instagood
#instadaily #ducati #white #motogp
#sportbike #sport #superbike #race
#racing #order #preorder

Figure 16: Sample spam tags

6 Question Extra Credit

6.1 Question

Work through the "Inductive SVM" example, discuss in detail the steps and resulting output

6.2 Methodology

Using the SVM^{light} which can be found [here](#), I was able to run the example and receive the results.

6.3 Results

Below is a screenshot of running SVM^{light} on the provided dataset as per the instructions.

Figure 17: Results from SVM^{light}

The program provided the scores of 96.43% for precision and 99% for recall. These are expected results since the example is created by the author of the program. The **svm_learn** executable goes through the training data and creates the support vectors, provides XiAlpha estimates and creates a model file to be used by the **svm_classify** executable.

The **svm_classify** executable uses the model file created by **svm_learn** to try to classify the test data into the classifications using the support vectors and provides the performance values which are very high for this set of test data. This means that the training data is very high quality.