Examining the cosine similarity of q_sentence embeddings in the English MP Corpus (based on all-mpnet-base-v2 embeddings)

We used the all-mpnet-base-v2 model to calculate sentence embeddings for all the quasi sentences. This model sits atop the leaderboard of sentence-transformer models for semantic search/comparison (https://www.sbert.net/docs/pretrained_models.html).

Due to computational and memory constraints, we calculated the cosine similarity between all q_sentence combinations, but only saved those combinations with a score above 0.8.

➔ In total, this gave us **379,546 combinations** of very similar quasi sentences

How similar are these sentences really? Do we need a cut-off higher than a cosine similarity of 0.8?

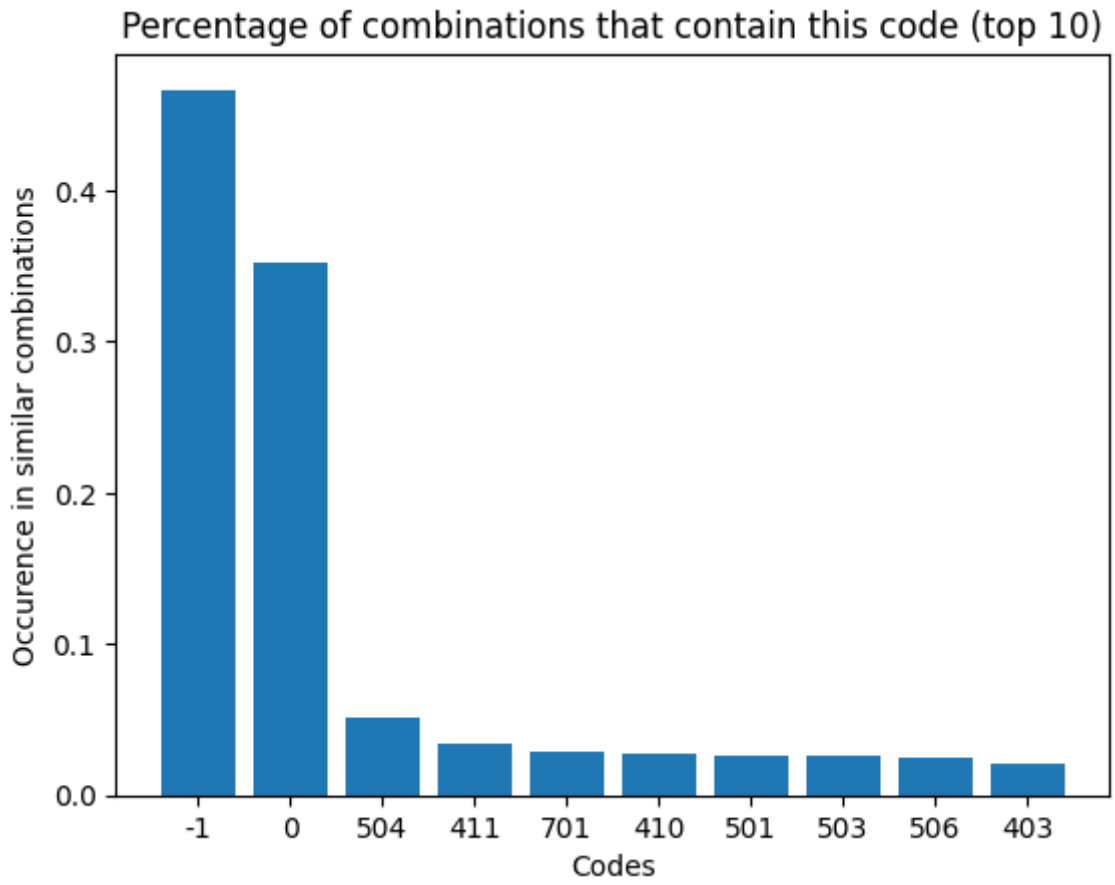
➔ Looking at the combinations with the lowest cosine similarity (so just above 0.8):
- "Give higher priority to the physical healthcare of those with mental health problems."

q_sentence 1	**q_sentence 2**	**Cosine similarity**
Give higher priority to the physical healthcare of those with mental health problems.	Prioritising mental healthcare	0.800000
and generate more local investment.	and stimulate their local economy.	0.800000
End exploitative zero-hours contracts.	banning exploitative zero-hours contracts	0.800002
Reducing unnecessary regulation.	Cut red tape and improved regulations	0.800003
every Australian has the right to access good quality health care, housing and income support.	Reasonable access to quality, affordable health care and education for all Australians, regardless of their location and personal circumstances.	0.800003
An Independent Scotland will	In an independent Scotland: we will always get the governments people vote for,	0.800004
Services may include:	The service needs the following:	0.800005
We must offer people with disabilities choice and control	We believe that disabled people should be supported and encouraged to follow	0.800005

	their aspirations, make their own choices, and to lead a quality life.	
In particular, tackling climate change is an economic necessity and the most important thing we must do for our children, our grandchildren and future generations.	We have a duty to future generations to protect our environment and tackle climate change.	0.800005
and open new markets for American goods and services in a competitive global economy.	and the opening of global markets for American business and exports to thrive.	0.800005
and the opening of global markets for American business and exports to thrive.	and open new markets for American goods and services in a competitive global economy.	0.800005
Family Friendly policy	We will promote family-friendly policies	0.800007

- ➔ These seem all very good, no higher threshold needed (we could even think of redoing the calculations with a lower threshold?
- ➔ ~ 76% of all combinations have the same code, so disagreement in ~24% of the time

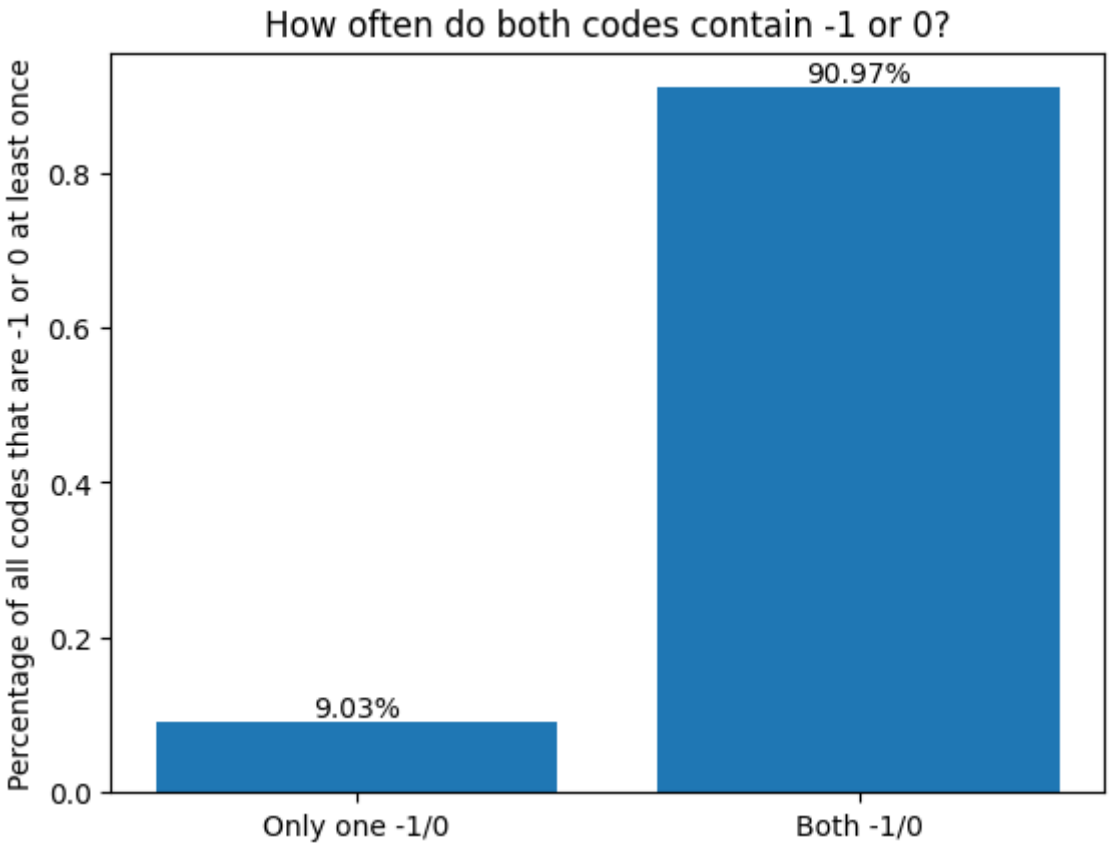
What codes are most common in these combinations?



➔ -1 corresponds to an NA code, 0 corresponds to “H” code (for Headers).

- So a large majority of the similar codes contain codes for “no meaning”

How often do both codes show either -1 or 0? So how often are both sentences coded as "no meaning" (when at least one contains -1 or 0)?

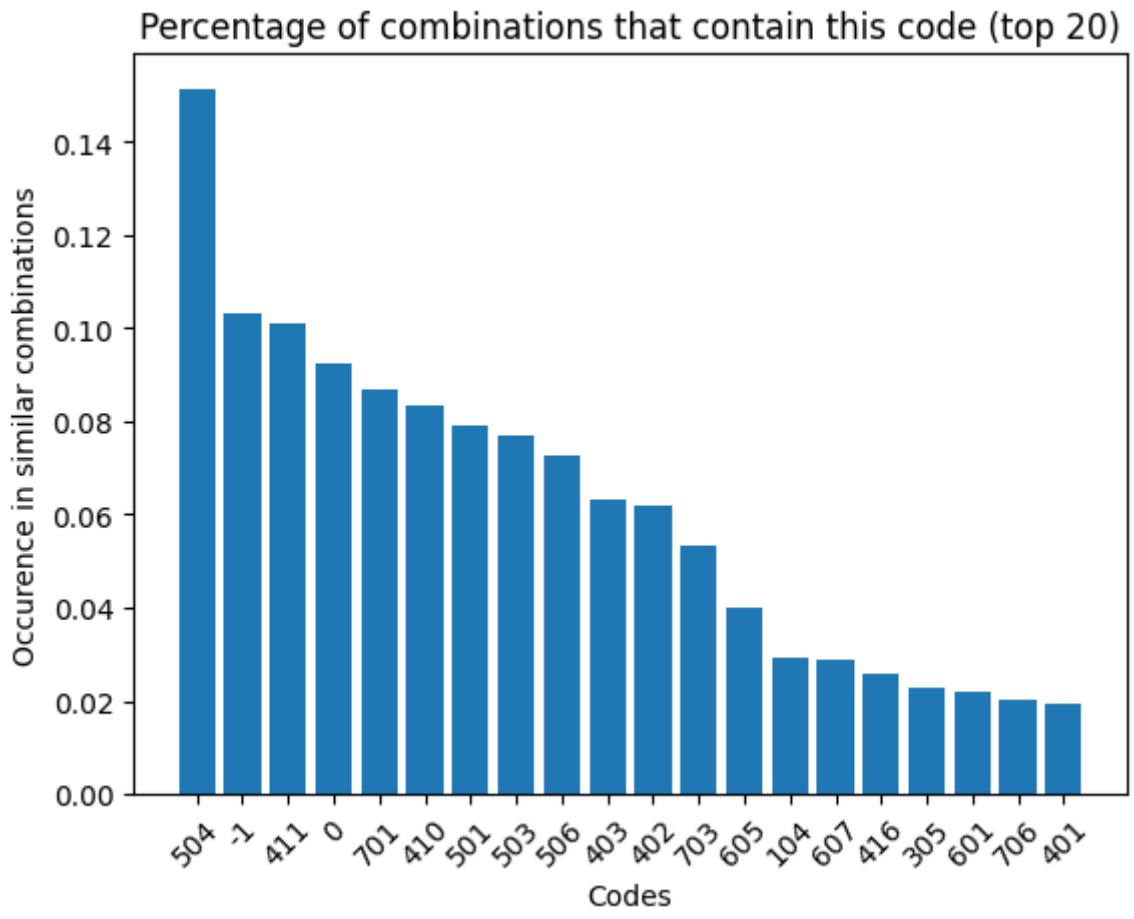


➔ So a large part of the combinations are two sentences that are very likely without meaning (header or fluff such as "We will:")
- These make up about 66% of all the combinations. **These are not as interesting to us, we will remove them and only look at combinations where at least one coder included a "real" code.**
- Still, in 9% of these codes (approx. 6.5% of all combinations) the coders do not agree whether these sentences contain real meaning or no meaning. That is quite a lot!

Here an example of some problematic codings based on this:

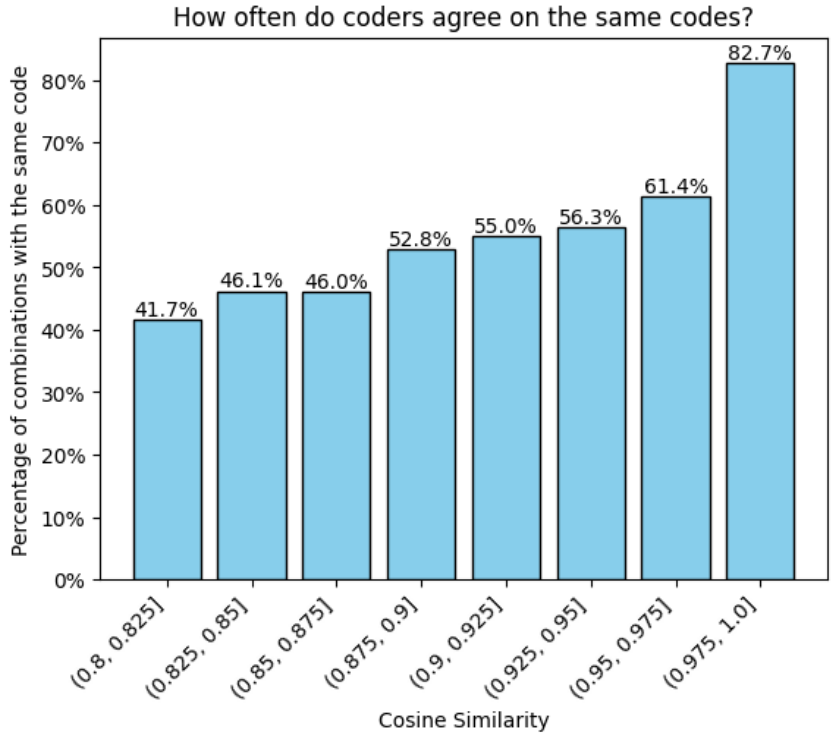
- Same sentence appears twice in the same document:
  - *"In the first year, we will double spending on books and equipment to overcome the effect of recent cuts."*
  - The **same coder** gave it **-1** code first, **506** (Education expansion) the second time
- Two different manifestos, two different coders.
  - Both manifestos contain the q_sentence "We will:"
    - One coder uses code -1 (makes sense). The other uses code 703 (Agriculture and Farmers: Positive). Likely the text afterwards talks about agriculture, but this just (wrongly) increases the number of occurrences of this code.

Now, after removing the “no meaning” combinations, what codes appear most often in a combination (**127,867 combinations remaining**)?

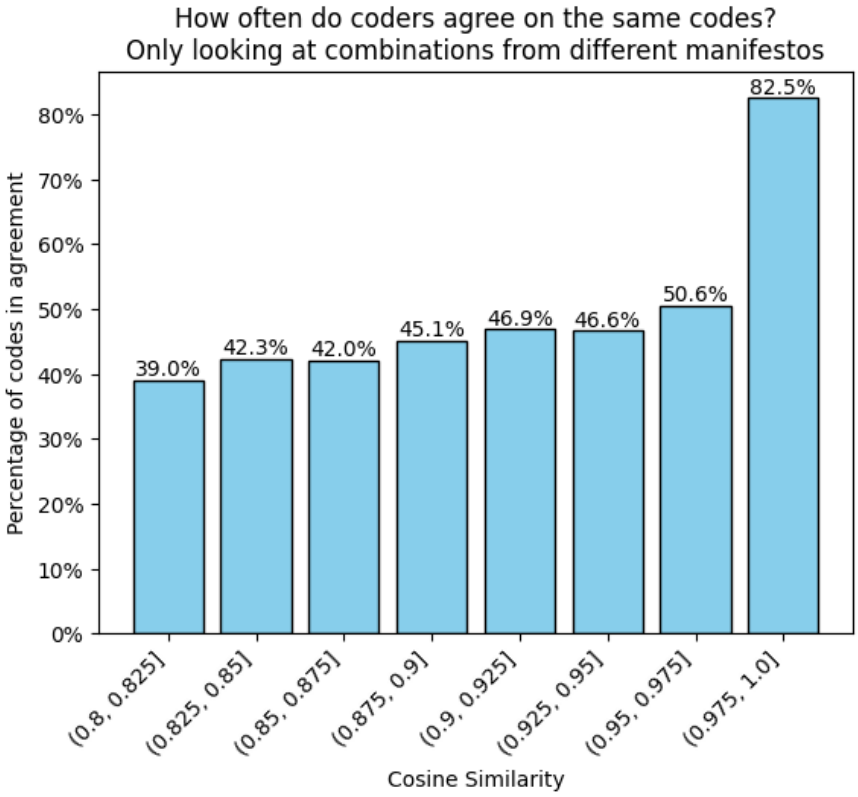


How often do coders agree on the same codes (having removed the "no meaning" combinations)?

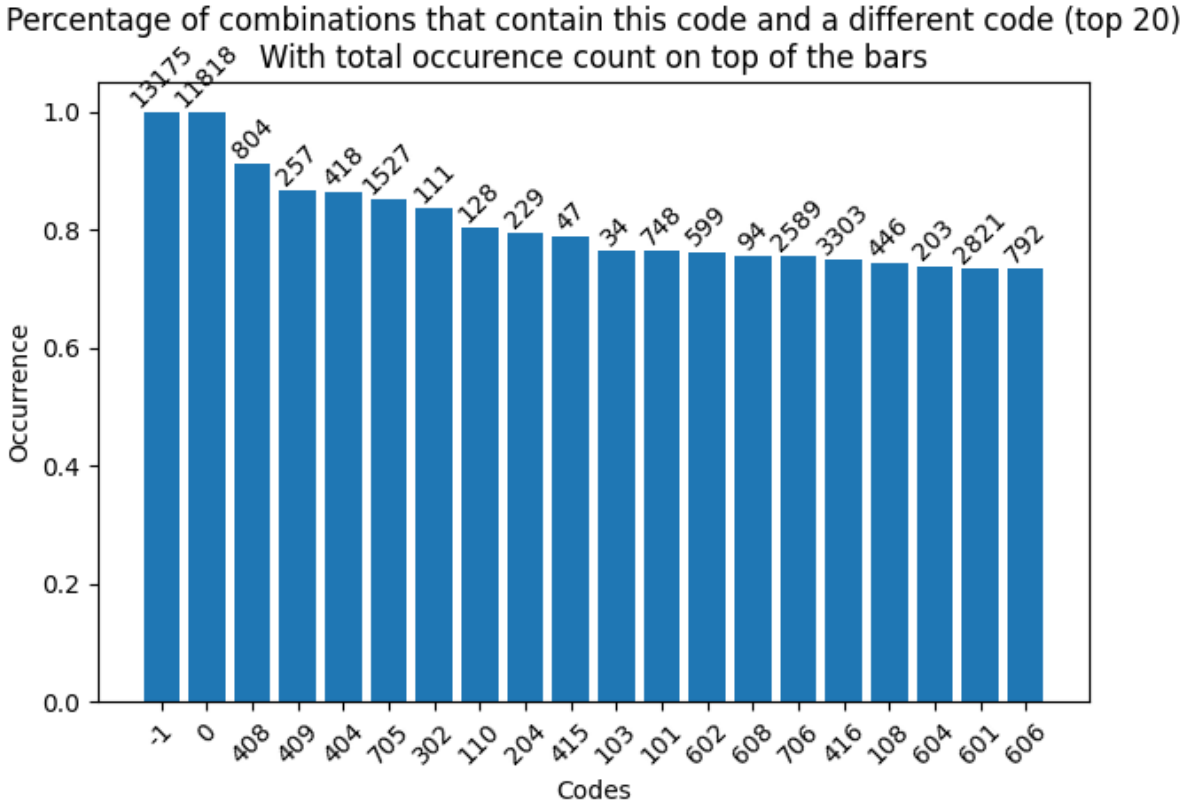
➔ On average, **44.5%** of the remaining combinations have different codes. What is the effect of higher cosine similarity on this?



The same plot, but removing all combinations from the same document (where a coder would likely just copy the same code)

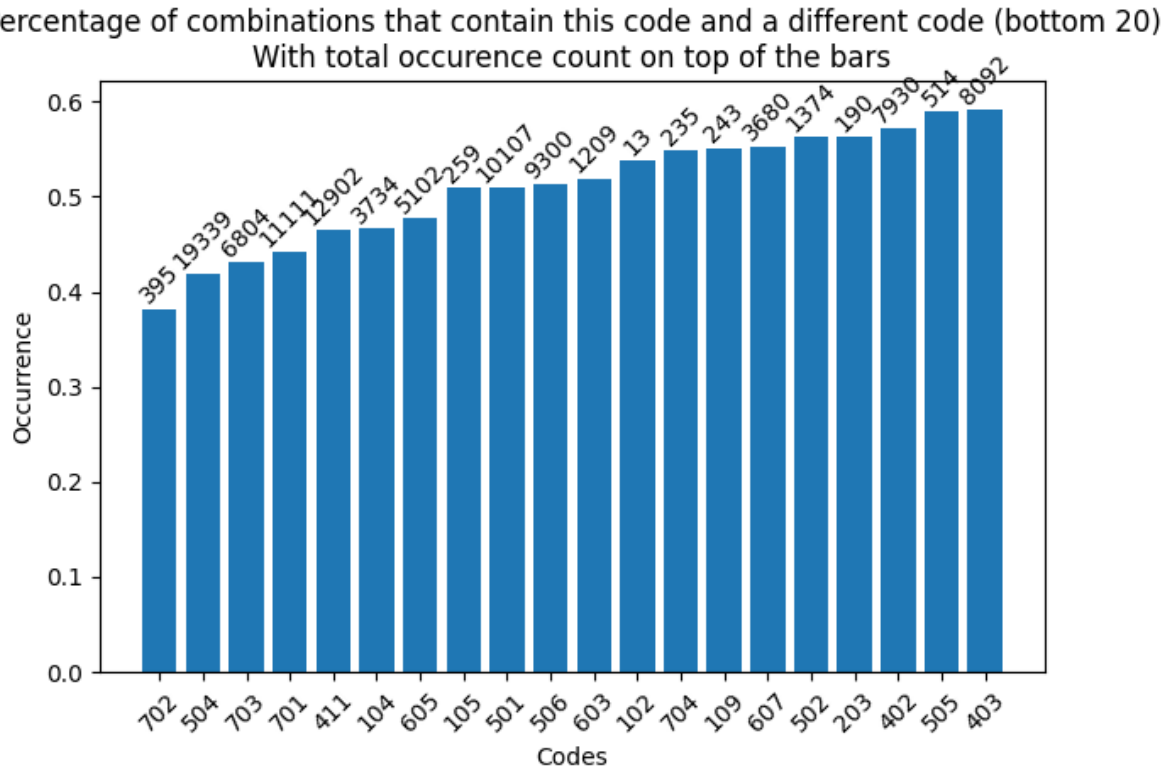


Which codes occur most often when the codes are not the same (most contentious codes?)



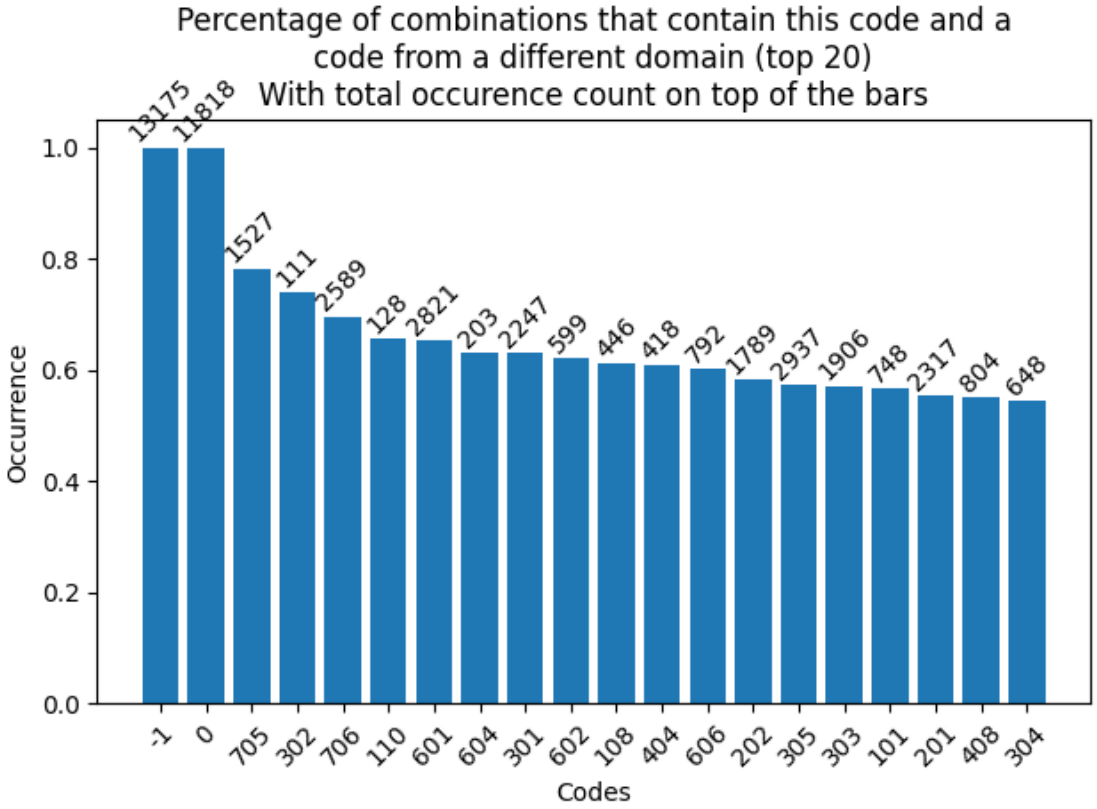
- That -1 and 0 are at 100% makes sense (we removed all where they were “in agreement”).
- 408: Economic Goals
    - Coding instruction: “Broad and general economic goals that are not mentioned in relation to any other category. General economic statements that fail to include any specific goal. Note: Specific policy positions overrule this category! If there is no specific policy position, however, this category applies.”
        - It seems that very often, a specific policy position can actually be found…
- 409: Keynesian Demand Management
- 404: Economic Planning
- 705: Underprivileged Minority Groups
- 302: Centralisation

Now the opposite, which codes are the least contentious (highest agreement when this code occurs)?

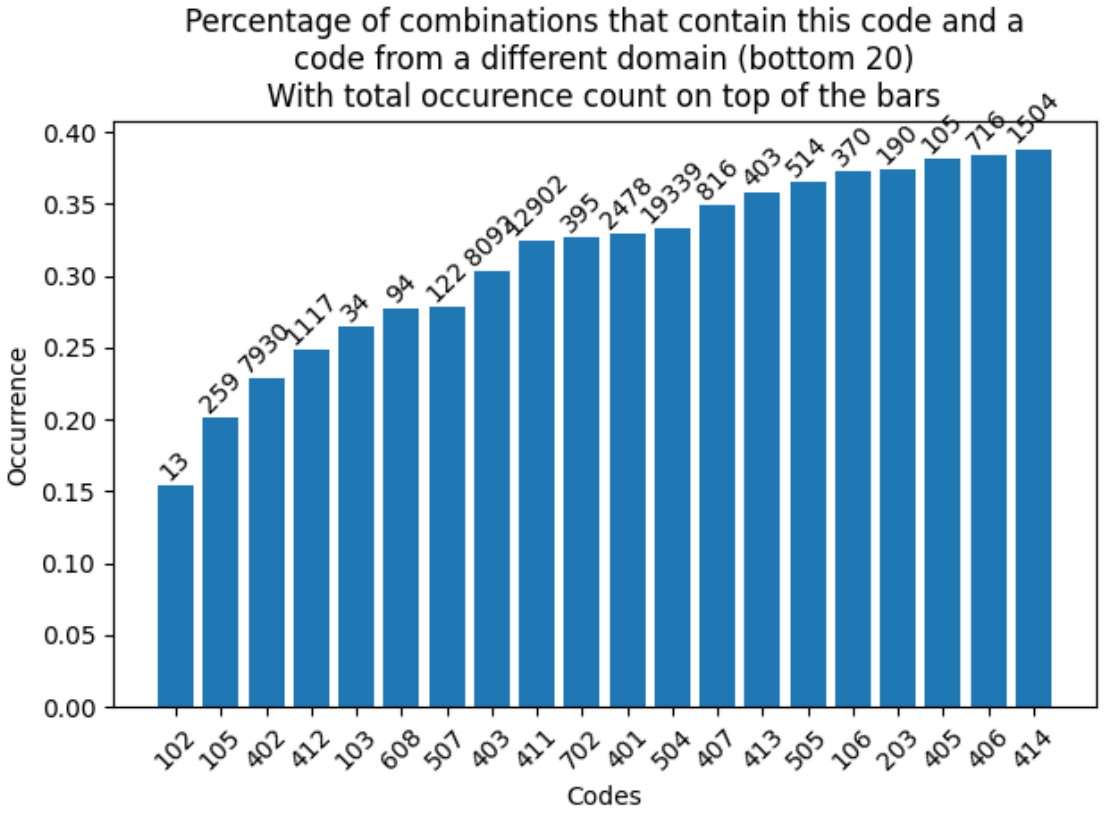


- 702: Labour Groups: Negative
- 504: Welfare State Expansion (remember, this was the most frequent real code)
- 703: Agriculture and Farmers: Positive
- 701: Labour Groups: Positive
- 411: Technology and Infrastructure

How often are codes very different (so from a different domain)? Most contentious codes:



And which codes are most consistent in regards to the domain?:



How often are the codes equal if the texts are completely the same?

- Approx. 89% of the combinations have the same code in this case
- Examples of exact same sentences with different codes:
    - “Together we move South Africa forward.”
        - 606 [Civic Mindedness: Positive] and 202 [Democracy]
    - “Education is key in helping girls and women escape the multi-generational poverty cycle.”
        - 503 [Equality: Positive] and 506 [Education Expansion]