(M) Adjupectiheaval! (1/3) [10 Points]

You are the administrator of the newest and greatest restaurant review site, whelp.com, which compiles reviews from the most noted gastronomical connoisseurs from around the world.

Recently, you've discovered that dishonest restaurants have been sneakily trying to increase their rating on Whelp! To do this, they're posting thousands of reviews written by spambots, small computer programs that pretend to be human reviewers. To ensure quality, you need to constantly delete these fake reviews. However, being just one administrator, you obviously can't read all of them manually.

Thankfully, spambots make some common mistakes in their fake reviews. Even if a review is grammatically correct, the review still might not make sense; some errors of this category can easily be spotted by anti-spam programs. For example, consider the following two reviews:

(A) At this restaurant, the cake is delicious yet satisfying.
(B) At this restaurant, the cake is delicious and satisfying.

One of these was probably written by a spambot, while the other could plausibly be a real review.

M1. Identify which sentence is spam.

The spam sentence is:

Sometimes, the mistakes made by a spambot may be more subtle. For example, the following sentence is quite reasonable:

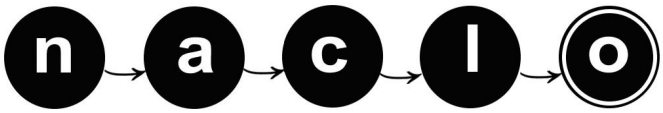
The cracker is crunchy and delicious.

But the following sentence is probably not written by a human (or, if so, one with bad taste):

The pudding is crunchy and delicious.

Of course, being able to make these judgements requires some knowledge of the foods involved.[1]

[1] More generally, this form of reasoning aided by human real-world knowledge is termed *knowledge-aware NLP*.



(M) Adjupectiheaval! (2/3)

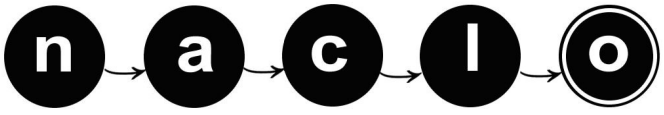
Having managed to filter out English-language spambots, you've decided to start investigating reviews in Bahasa Indonesia, the national language of Indonesia. However, your task is complicated by the fact that you don't speak Indonesian! In order to write filtering software, you first examine some reviews written by real humans, about popular Indonesian foods such as *kemplang* and *poffertjes*.

1. *Kemplang manis namun berminyak.*
2. *Rengginang manis dan lezat.*
3. *Poffertjes manis serta lezat.*
4. *Rempeyek lezat dan menggugah selera.*
5. *Lemang menggugah selera dan manis.*
6. *Onde-onde lezat namun mahal.*
7. *Poffertjes baik namun mahal.*
8. *Kemplang baik dan sehat.*
9. *Lemang sehat serta manis.*
10. *Rempeyek berminyak dan hambar.*
11. *Rengginang tidak sehat serta mahal.*
12. *Onde-onde berminyak dan tidak sehat.*

Despite not knowing anything about the food items mentioned in the reviews, or anything about the Indonesian language itself, you realize that this is enough to filter out some spam reviews!

M2. Below are six reviews. Three of them are almost certainly spam, while the other three could have been written by a human. Indicate whether the review is real or spam, where "real" means it could be a real review and "spam" means it's probably spam.

13. *Kemplang menggugah selera serta baik*. real / spam
14. *Rengginang hambar namun sehat*. real / spam
15. *Poffertjes baik namun tidak berminyak*. real / spam
16. *Rempeyek tidak manis serta berminyak*. real / spam
17. *Lemang manis dan sehat*. real / spam
18. *Onde-onde sehat namun tidak menggugah selera*. real / spam



(M) Adjupectiheaval! (3/3)

The algorithm you've designed using this knowledge works well, but you find that there are still some words and reviews that stump it. Here are some examples of real (non-spam) sentences in Indonesian that confuse your algorithm:

19. *Onde-onde halus dan manis.*
20. *Rengginang halus serta hambar.*
21. *Rempeyek garing serta baik.*
22. *Lemang tidak mahal namun garing.*
23. *Lemang halus dan tidak mahal.*
24. *Onde-onde berminyak dan garing.*
25. *Renggingang lembut namun lezat.*
26. *Lemang lembut namun mahal.*
27. *Rempeyek garing dan sehat.*

You quickly realize that to fully understand these sentences, you're going to have to read up more about these food items. Unfortunately, you only have access to a monolingual Indonesian dictionary (entries below): [2]

- ***Kemplang*** *adalah sebuah kerupuk ikan yang umum ditemukan di belahan selatan Sumatra, Indonesia. Kerupuk kemplang dikeringkan dan kemudian dipanggang atau digoreng.*
- ***Lemang*** *adalah kue dari beras ketan yang dimasak dalam seruas bambu, setelah sebelumnya digulung dengan selembar daun pisang.*
- ***Rempeyek*** *adalah sejenis makanan pelengkap dari kelompok gorengan. Fungsi rempeyek sama dengan kerupuk yaitu sebagai pelengkap hidangan.*
- ***Rengginang*** *adalah sejenis kerupuk tebal yang terbuat dari beras ketan dibentuk bulat yang digoreng panas dalam minyak goreng.*
- ***Onde-onde*** *adalah sejenis kue yang populer di Indonesia. Ini sangat terkenal di daerah Mojokerto yang disebut sebagai kota onde-onde sejak zaman Majapahit.*
- ***Poffertjes*** *adalah kue tradisional yang empuk dari Belanda. Penampilannya mirip panekuk, tetapi lebih kecil dan manis.*

With this new information, you find that you can deduce which reviews are real or spam!

M3. For each of reviews 28-31, indicate whether the review is real or spam, where "real" means it could be a real review and "spam" means it's probably spam.

28. *Onde-onde halus serta mahal.* real / spam
29. *Rempeyek lembut namun tidak sehat.* real / spam
30. *Kemplang garing dan tidak berminyak.* real / spam
31. *Poffertjes garing serta hambar.* real / spam

[2] Adapted from Bahasa Indonesia Wikipedia.