SAE S1.01
Projet "chatbot" - Partie 1

Lisez d'abord attentivement les sections 1, 2 et 3 pour comprendre l'objectif du projet. Vous serez guidé pour atteindre cet objectif dans les sections suivantes. Des informations importantes sur les aspects organisationnels du projet sont donnés dans la section 4. Enfin, cette première partie pose les bases d'un programme qui sera amélioré dans une seconde partie.

1. Tâche

L'objectif du projet est de créer un chatbot aussi fiable et aussi rapide que possible pour répondre à des questions de culture générale. Ci-dessous, un exemple de dialogue qui devrait être possible à la fin du projet (le chatbot en gras).

> J'attends tes questions de culture générale.
> Qui a inventé le téléphone ?
> Alexander Graham Bell a inventé le téléphone.
> Qui a peint « La Joconde » ?
> La Joconde a été peinte par Léonard de Vinci.
> En quelle année ?
> La Joconde a vraisemblablement été peinte entre 1503 et 1507.
> Quelle est la capitale du Brésil ?
> Je ne sais pas.
> Je vais te l'apprendre.
> Je t'écoute.
> La capitale du Brésil est Brasilia.
> Très bien, c'est noté.
> Quelle est la capitale du Brésil ?
> **La capitale du Brésil est Brasilia.**

On remarque à travers cet exemple de dialogue que le chatbot répond correctement aux questions y compris lorsqu'elles supposent la prise en compte d'un contexte *(« En quelle année ? »* se réfère à la question/réponse précédente). On remarque aussi que la réponse n'est pas stéréotypée (la réponse à la question *« Qui a … ? »* prend des formes variées). On peut aussi noter que le chatbot répond *« Je ne sais pas. »* lorsqu'il n'a pas la réponse. Enfin, on observe qu'il est capable d'intégrer les connaissances transmises par l'utilisateur (la réponse concernant le nom de la capitale du Brésil dans l'exemple).

2. Principes de fonctionnement du programme

Le chatbot va répondre à la question de l'utilisateur en choisissant une réponse dans un ensemble de réponses données (dans la suite, on parlera de la base de réponses). Il va procéder en **2 étapes**. La **première étape** consiste à **rechercher les réponses qui sont liées au thème** de la question. Par exemple, si la question porte sur l'invention du téléphone, les réponses *« Le téléphone a été inventé par Alexander Graham Bell. »* et *« Le téléphone a été inventé en 1876. »* sont retenues. La **seconde étape** consiste à **sélectionner** parmi ces réponses retenues, celles dont la **forme est cohérente avec la question**. Par exemple, si la question est *« Qui a inventé le téléphone ?»*, une réponse de forme *« Le … a été … par … »* est possible. Alors que si la question est *« Quand le téléphone a-t-il été inventé ? »*, la forme de réponse attendue est plutôt : *« Le … a été … en … ».*

2.1 Etape 1 : Recherche des réponses candidates basée sur le thème abordé par la question

Afin de sélectionner dans notre base de réponses, celles correspondant au thème de la question, nous nous appuierons sur les mots de la question **exceptés les mots-outils** (« le », « a », « été », « par », « en », …) qui n'apportent pas d'information sur la thématique. Nous rechercherons ainsi les réponses qui contiennent l'intégralité des mots non outils de la question, qu'on appellera **réponses candidates**. Par exemple, pour la question *« Qui a inventé le téléphone ? »* ou *« Quand a été inventé le téléphone ? »*, nous rechercherons les réponses contenant les mots *« inventé »* et *« téléphone »*.

Afin de trouver ces réponses, une approche naïve consisterait à parcourir, à chaque nouvelle question, l'ensemble de la base des réponses à la recherche des réponses contenant tous les mots non outils de la question. Cette approche serait

très coûteuse en temps de calcul et les délais de réponse très longs pour des bases de réponses de grande taille. Une meilleure solution consiste, préalablement à toute question de l'utilisateur, à construire un **index**, c'est-à-dire une table répertoriant pour chaque mot non outil (entrée de l'index), les réponses (identifiées par un numéro) qui le contiennent. Il suffira alors, pour construire la liste des réponses candidates de faire l'intersection des réponses trouvées directement dans l'index pour chaque mot non outil de la question.

ILLUSTRATION DU FONCTIONNEMENT EN ENTONNOIR DU CHATBOT

QUESTION : « Qui a inventé le téléphone ? »

1- **ETAPE 1** : RECHERCHE DES REPONSES CANDIDATES SUR LA BASE DU THEME
basée sur les mots non outils de la question « Qui a inventé le téléphone ? »
Résultat :
→« Le téléphone a été inventé par Alexander Graham Bell. »
→« Le téléphone a été inventé en 1876. »
→« Alexander Graham Bell a inventé le téléphone. »
→« Le téléphone a été inventé aux Etats-Unis. »

2- **ETAPE 2** : SELECTION PARMI LES REPONSES CANDIDATES SUR LA BASE DE LA FORME
basée sur les mots-outils de la question « Qui a inventé le téléphone ? »
Résultat :
→« Le téléphone a été inventé par Alexander Graham Bell. »
→« Alexander Graham Bell a inventé le téléphone. »

3- CHOIX AU HASARD D'UNE REPONSE DANS LA SELECTION ISSUE DE L'ETAPE 2
REPONSE : « Le téléphone a été inventé par Alexander Graham Bell. »

2.2 Etape 2 : Sélection (parmi les réponses candidates) des réponses dont la forme est cohérente avec la question

Une fois les réponses sélectionnées sur la base du thème de la question, on dispose après l'étape 1, d'un ensemble de réponses candidates. Par exemple, si la question est *« Qui a inventé le téléphone ? »*, la réponse *« Le téléphone a été inventé par Alexander Graham Bell. »* et la réponse *« Le téléphone a été inventé en 1876. »* sont candidates puisqu'elles contiennent toutes les deux l'intégralité des mots non outils (*« téléphone »* et *« inventé »*) de la question. L'objectif dans l'étape 2 est de sélectionner parmi ces réponses candidates, celles dont la forme est cohérente avec la forme de la question. Pour définir les formes, on ne s'appuiera cette fois-ci **que sur les mots-outils**.

Afin d'établir la cohérence des formes des réponses avec celles des questions, on dispose d'un ensemble de **questions/réponses idéales**. Il est possible de chercher dans cet ensemble, les formes de réponses possibles étant donnée la forme de la question. Par exemple, si dans cette base de question/réponses idéales, on trouve les 2 questions/réponses *« Qui a découvert le continent américain ? Le continent américain a été découvert par Christophe Colomb. »* et *« Qui a inventé le stéthoscope ? René Laennec a inventé le stéthoscope. »*, on en déduira que des réponses de forme *« Le ... a été ... par ... »* et *« ... a ... le ... »* sont possibles pour une question de forme *« Qui a ... le ... ? »*.

Pour cette recherche des formes possibles, une approche naïve consisterait à parcourir, à chaque question, l'ensemble des questions/réponses idéales. Cette approche serait trop coûteuse en temps de calcul. Nous proposons donc, comme dans l'étape 1, de créer un index afin d'accélérer cette recherche. Dans cet index, les formes de réponses possibles identifiées par un numéro seront répertoriées pour les mots outils des questions (entrées de l'index). De plus, afin de distinguer des questions dont la forme ne varierait que sur la position des mots-outils mais dont la forme de la réponse serait différente, nous allons considérer la position du mot-outil par rapport aux autres mots-outils de la question. Cela se traduira dans l'index par une entrée correspondant à une concaténation du mot-outil avec son indice (par exemple *« quel_1 »*, pour la question *« Dans quel pays est située Paris ? »* et « *quel_0 »* pour la question *« Quel est le personnage principal dans le roman « notre Dame de Paris » ? »*).

Après avoir déterminé les formes de réponses possibles pour la question grâce à l'index, il suffira de choisir parmi les réponses candidates issues de l'étape 1, celles qui correspondent à une des formes possibles. Dans notre exemple, si la question est *« Qui a inventé le téléphone ? »*, les réponses *« Le téléphone a été inventé par Alexander Graham Bell. »* ou *« Alexander Graham Bell a inventé le téléphone. »* seront considérées comme cohérentes puisqu'elles correspondent chacune à une des formes possibles listées ci-dessus. Au contraire, la réponse candidate *« Le téléphone a été inventé en 1876. »* sera considérée comme non cohérente puisqu'elle ne correspond à aucune forme de réponse possible pour cette question. On peut remarquer qu'avec ce principe de fonctionnement, la question/réponse exacte n'a pas besoin d'être présente dans notre base de questions/réponses idéales, il suffit qu'une autre question/réponse de forme similaire le soit.

Lorsque plusieurs réponses sont sélectionnées à l'issue de l'étape 2, on en choisira une au hasard. Cela aura l'avantage de donner une certaine variabilité aux réponses du chatbot et un caractère plus naturel à l'échange.

3. Les fichiers et leurs formats

Les données nécessaires au fonctionnement de votre chatbot sont dans des fichiers ayant un format bien défini décrit ci-dessous. Libre à vous d'en modifier le contenu en veillant à en conserver le format.

3.1 Le fichier des réponses

Les réponses sont regroupées dans le fichier ***reponses.txt***. Une réponse est une phrase simple. Un seul élément d'information est donné par réponse. Il y a par exemple une réponse qui donne le nom de l'auteur de l'invention et une réponse qui en donne l'année.

```
Les frères Lumière ont inventé le cinéma.
Auguste a fondé l'Empire romain.
Cai Lun a inventé le papier.
Isaac Newton a découvert les lois du mouvement.
Pépin le Bref a fondé la dynastie des Carolingiens.
René Laennec a inventé le stéthoscope.
Le microscope a été inventé en 1670.
Le papier a été inventé vers 105.
Le stéthoscope a été inventé en 1816.
Antonie van Leeuwenhoek a inventé le microscope.
Sigmund Freud a fondé la psychanalyse.
Homère a écrit L'Odyssée
…
```

3.2 Le fichier des questions-réponses idéales

Les questions/réponses idéales sont regroupées dans le fichier ***questions-reponses.txt***. Il y a une question/réponse par ligne. La réponse est séparée de la question par « ? ».

```
Qui a inventé le cinéma ? Les frères Lumière ont inventé le cinéma.
Qui a écrit L'Alchimiste ? L'Alchimiste a été écrit par Paulo Coelho.
Quel pays est la patrie des samouraïs ? La patrie des samouraïs est le Japon.
Dans quel pays est située Caracas ? Caracas est située au Venezuela.
Qui a découvert Pluton ? Clyde Tombaugh a découvert Pluton.
Quelle est la capitale du Laos ? La capitale du Laos est Vientiane.
En quelle année le Titanic a-t-il coulé ? Le Titanic a coulé en 1912.
Quel est le surnom de la Joconde ? La Joconde est surnommée Mona Lisa.
…
```

3.3 Le fichier des mots-outils

Les mots-outils sur lesquels les calculs de forme sont faits sont regroupés dans le fichier ***mots-outils.txt***. Il y a un mot par ligne.

```
à
a
afin
après
alors
année
…
```

4. Organisation du travail

Le projet sera réalisé en binôme (défini par vos enseignants) et donnera lieu à la rédaction d'un rapport et d'une présentation orale lors de la dernière séance. En plus de votre rapport, il vous faudra rendre les fichiers correspondant à vos programmes ainsi que les fichiers `reponses.txt`, `questions-reponses.txt` et `mots-outils.txt`. Commencez par lancer le script ***debut-sae-s1-01*** permettant la récupération du répertoire ***projet-sae-s1-01*** contenant le projet IntelliJ nommé `ProjetChatbot`. Ouvrez-le. La procédure `main` est dans la classe `Chatbot`. En l'état, elle répond aléatoirement aux questions des utilisateurs. Testez-la. Parcourez les fichiers sources récupérés pour découvrir les différentes classes, structures de données et les différentes méthodes mises à votre disposition.

5. Développements des étapes 1 & 2 du choix de la réponse du chatbot

Dans cette première partie, nous vous guidons pour réaliser les étapes 1 & 2 décrites ci-dessus. A l'issue de ces premiers développements, bien qu'imparfait, votre chatbot devrait être en mesure de répondre rapidement et correctement à bon nombre de questions (encore faut-il que la réponse soit présente dans le fichier des réponses et que celle-ci contienne tous les mots non outils de la question). **Les spécifications détaillées de chaque méthode sont données dans le code source fourni.**

5.1 Développement de la classe Index

a) Les index sur le contenu et la forme vont nous permettre de retrouver respectivement les réponses candidates et les formes de réponses appropriées étant donnée la question. En analysant le code de la classe Index, faites un dessin représentant votre compréhension du contenu de :

- *l'index sur le contenu basé sur* `mini_reponses.txt` *sachant qu'une réponse est identifiée par un entier correspondant à son indice dans la table des réponses et que cette table respecte l'ordre* ***du fichier des réponses****.*
- *l'index sur la forme basé sur* `mini_questions-reponses.txt` *sachant qu'une forme est identifiée par un entier correspondant à son indice dans la table des formes et que cette table respecte l'ordre dans laquelle on trouve cette forme dans* ***le fichier des réponses****.*

b) Dans **Index.java***, développez les méthodes* `rechercherSortie` *et* `ajouterSortie` *de la classe* `EntreeIndex`.

c) Dans **Index.java**, *développez les méthodes* `rechercherEntree`, `ajouterSortieAEntree` *et* `rechercherSorties` *de la classe* `Index`.

5.2 Construction et utilisation d'un index pour construire la liste des réponses candidates (étape 1)

a) Dans **Utilitaire.java**, *développez les méthodes* `trierChaines` *et* `existeChaineDicho` *pour trier le vecteur des mots-outils et vérifier par recherche dichotomique la présence d'un mot dans ce vecteur.*

b) Dans **Utilitaire.java**, *développez la méthode* `constructionIndexReponses` *qui construit l'index dont l'entrée correspond aux mots (non outils) des réponses et les sorties correspondent aux identifiants des réponses (les indices dans la table des réponses) contenant ces mots.*

c) Dans **Chatbot.java**, *dans la méthode* `main` *initialiser et trier* `motsOutils` *et appelez la méthode* `constructionIndexReponses` *pour initialiser* `indexThemes`. *Affichez l'index construit pour* `mini_reponses.txt` *pour vérifier le bon fonctionnement de la méthode.*

d) Dans **Utilitaire.java**, *développez les méthodes* `fusion`, `maxOccurences` *et* `constructionReponsesCandidates` *pour construire l'ensemble des réponses candidates.*

e) *Dans* **Chatbot.java**, *dans la méthode* `repondre` *appelez la méthode* `constructionReponsesCandidates` *pour initialiser* `reponsesCandidates`. *Remplacez l'affichage aléatoire par l'affichage des réponses candidates. Si* `reponsesCandidates` *est vide, affichez* `« Je ne sais pas. »`

5.3 Construction et utilisation d'un index pour sélectionner parmi les réponses candidates celles dont la forme est cohérente avec celle de la question (étape 2)

a) Dans **Utilitaire.java**, *développez les méthodes* `calculForme`, `rechercherChaine`, `existeChaine` et `constructionTableFormes` *pour construire la table des formes de réponses possibles. Appelez* `constructionTableFormes` *dans la procédure* `main` *de* **Chatbot.java** *pour initialiser* `formesReponses`.

b) Dans **Utilitaire.java**, *développez la méthode* `constructionIndexForme` *qui construit l'index dont l'entrée correspond aux mots-outils positionnés des questions (par exemple « qui_0 ») et les sorties aux identifiants des formes de réponses (indice dans la table des formes) possibles. Appelez* `constructionIndexForme` *dans la procédure* `main` *de* **Chatbot.java** *pour initialiser* `indexFormes`. *Affichez l'index construit pour* `mini_questions-reponses.txt` *pour vérifier.*

c) Dans **Utilitaire.java**, *développez la méthode* `selectionReponsesCandidates` *qui sélectionne parmi les réponses candidates celles dont la forme est cohérente avec la question.*

d) Dans **Chatbot.java**, *dans la méthode* `repondre`, *dans le cas où* `reponsesCandidates` *est non vide, appelez la méthode* `selectionReponsesCandidates` *pour initialiser* `reponsesSelectionnees`. *Faites afficher les réponses sélectionnées pour vérifier le bon fonctionnement de la méthode. Dans le cas où* `reponsesSelectionnees` *est vide, affichez* « `Je ne sais pas.` »

e) Dans **Chatbot.java**, *commentez tous les affichages intermédiaires, sélectionnez au hasard une réponse de* `reponsesSelectionnees` *et affichez-la pour produire la réponse du chatbot.*

Un exemple de trace que vous devriez pouvoir produire à ce stade :

```
> J'attends tes questions de culture générale.
> Qui a inventé le téléphone ?
> Le téléphone a été inventé par Alexander Graham Bell.
> En quelle année le téléphone a-t-il été inventé ?
> Le téléphone a été inventé en 1876.
> Quelle est la capitale de la France ?
> La capitale de la France est Paris.
> Quelle est la capitale de l'Italie ?
> La capitale de l'Italie est Rome.
> Quelle est la capitale de la Crète ?
> Je ne sais pas.
> Au revoir
```

Bien que maintenant fonctionnel, notre chatbot reste très perfectible. A la question « `Quel est le surnom de la Joconde ?` », la réponse est « `Je ne sais pas.` » car bien que la réponse « `La Joconde est surnommée Mona Lisa.` » soit présente dans notre base de réponses, elle n'est pas retenue comme réponse candidate puisque le mot « `surnom` » n'est pas présent dans cette réponse (`surnommée` ≠ `surnom`).

Il est par ailleurs fastidieux de formuler intégralement une question quand on demande juste une précision. Dans la trace ci-dessus, on aimerait pouvoir simplement écrire « `En quelle année ?` » ou « `Quand ?` » plutôt que « `En quelle année le téléphone a-t-il été inventé ?` ».

Il serait aussi souhaitable de pouvoir transmettre de nouvelles connaissances à notre chatbot de façon interactive sans avoir à éditer directement les fichiers `reponses.txt` et `questions-reponses.txt`.

Il s'agit d'une partie des améliorations que nous nous proposons d'apporter dans la seconde partie que nous vous invitons à démarrer dès la première partie terminée.